MSDS410 RegressionAnalysis and Multivariate Methods Fall 2018

Instructor
Melvin Ott, PhD
melvin.ott@northwestern.edu

Welcome to MSDS 410. I hope you find this course to be challenging and rewarding. Along with the challenge, the assignments should be realistic and enjoyable. Plan your time so you can do some research on the topics in the major assignments and to provide solutions that are thoughtful and creative. I will reward creativity, especially if it works.

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 Python statistical software and its use in data management and statistical modeling.

Prerequisites: MSDS 400 Math for Data Scientists and MSDS 401 Statistical Analysis.

Required Text

Chatterjee, S. and Hadi, A. S. 2012. *Regression Analysis by Example*, fifth edition. New York, NY: Wiley [ISBN-13: 978-0470905845]

The required text should be available at www.abbotthall.bncollege.com

Optional Text (if you need it)

VanderPlas, J., 2017. *Python Data Science Handbook: Essential Tools for Working with Data*. Sebastopol, Calif.: O'Reilly [ISBN-13: 978-1491912058]

Optional text available at many online locations.

Software

No software purchases are needed for this course. Software in this course is freely available on the web for PC/Windows and Mac/OS X systems. The primary software environment for this course is Python as implemented in Anaconda, Enthought Canopy platform or Rodeo.

Recommend you install Python 3.6 from https://www.anaconda.com/download/ this will give you both Spyder and Jupyter. See the Anaconda CheatSheet in course downloads for more info.

There are many online resources for Python training. Here are a few good books for learning more about Python:

Beazley, D. M. 2009. *Python Essential Reference* (4th ed.). Boston: Addison-Wesley. [978-0672329784]

Beazley, D. M. 2017. *Python Programming Language Live Lessions*, Old Tappan, N.J.: Pearson.Video available through Safari Books Online, Sebastopol, Calif.: O'Reilly.

Beazley, D. & Jones, B. K. 2013.*Python Cookbook* (3rded.). Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-1-449-34037-7]

Chun, W. J. 2007. *Core Python Programming* (2nded.). Upper Saddle River, N.J.: Prentice Hall. [ISBN-13: 978-0-13-226993-3]

Gift, N. and Jones, J. M. 2008.*Python for Unix and Linux System Administrators: Efficient Problem Solving with Python.* Sebastopol, Calif.: O'Reilly. (Chapter 2: IPython, pages 21–69.) [ISBN-13: 978-0-596-51582-9]

Hellmann, D. 2011. *The Python Standard Library by Example.* Upper Saddle River, N.J.: Pearson/Addison-Wesley. [ISBN-13: 978-0-321-76734-9]

Lubanovic, B. 2015.*Introducing Python: Modern Computing in Simple Packages.* Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-1-449-35936-2]

Ramalho, L. 2015. *Fluent Python: Clear, Concise, and Effective Programming.* Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-1-491-94600-8]

Sweigart, A. 2015.*Automate the Boring Stuff with Python: Practical Programming for Total Beginners.* San Francisco: No Starch Press. [978-1-59327-599-0]

A useful overview of the world of Python, *The Hitchhiker's Guide to Python* by Kenneth Reitz, is available online at<http://docs.python-guide.org/en/latest/>.

Stephenson, B. 2015.*The Python Workbook: A Brief Introduction with Exercises and Solutions.* New York: Springer [ISBN-13: 978-3-319-14239-5] Available to Northwestern University students as a free downloaded from the Springer collection at http://link.springer.com.turing.library.northwestern.edu/.

Lynda.com offers a wide range of technical training materials, and Northwestern University has a contract with Lynda.com to provide these courses at no cost to registered students. The program is described at:

http://www.northwestern.edu/hr/workplace-learning/lynda

Of special interest to students in this course are Lynda.com courses in Python and Git. The Lynda.com system includes many Python courses. There is also a one-hour and twenty-minute lecture titled *Up and Running with Git and GitHub.* Another Lynda.com course *Git Essential Training* is a more complete six-hour course on Git technology.

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.
- 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.

Evaluation

Final grade will be determined from a total of 1000 possible points as follows:

Bonus Points	(100+ possible points)
Assignments	(900 possible points from homework assignments)
Quizzes	(100+ possible points)

Assignments:MSDS 410 will have three graded assignments. Assignment #1 and #2 will each have 300 possible points. Assignment 4A will have 100 points. Assignments #3 & #4 are tutorial assignments to let you have some hands on Python experience. You will receive 100 points each for completing and turning in assignments #3 and #4.

Grading Scale

A	**= 930–1000 points**
A-	**= 900–929 points**
B+	**= 870–899 points**
B	**= 830–869 points**
B-	**= 800–829 points**
C+	**= 770–799 points**
C	**= 730–769 points**
C-	**= 700–729 points**
F	**= 000–699 points**

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.

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.

Attendance

This course will not meet at a particular time each week. All course goals, learning objectives, and assessments are supported through classroom elements that can be accessed at any time. 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.

Learning Groups

Student study groups will be available in this course as a means to foster a collaborative learning environment. Blue Jeans is available as a conferencing tool.

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 School of Professional Studies student handbook at <http://sps.northwestern.edu/program-areas/graduate/student-handbook.php> 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.

All courses operate on a Monday to Sunday schedule.

Week 1 Topic: Exploratory Data Analysis and Simple Linear Regression

Learning Objectives

After this week, you should be able to:

- Explain the importance and the role of exploratory data analysis in statistical modeling.
- 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

Your tour guide has provided a focused reading approach to these topics.
Regression Analysis by Example Chapters 1–2,pages 1–49

Note p 27 re covariance and standardize data, p28 correlation, p29 Anscombe Quartet, p 32 simple linear regression, p 33 parameter estimation, p36-40 hypothesis tests, p41-43 predictions, p43-46 quality of fit and r squared and p46 regression through the origin.

Optional Reading (if you need it)

Python Data Science Handbook, Chapter 4 pages 217-238, 311-330
Note p 217-238 plotting and plot options, p 311-330 plotting with Seaborn.

Discussion Board

Join in the week-specific discussion board forum. Your participation in both posting and responding will help you gain a better understanding of the material and provide interaction with other students.

Assignments

No assignment due for week one.

Sync Session

The date of the first sync session will be announced.
This sync session will cover:

- Course overview
- Q&A
- Linear Regression

It will be recorded and posted for the benefit of students who are unable to attend. You are encouraged to attend this sync session or view the sync session recording. If you would like to ask questions you have to be there.

Week 2 – Topic: Multiple Linear Regression

Learning Objectives

After this week, the student will be able to:

- Fit and interpret a multiple linear regression model.
- Compute and interpret the statistical tests associated with multiple linear regression .
- Understand the analysis of variance table and the associated metrics and tests of significance for multiple linear regression.
- 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 multiple linear regression.

Assigned Reading

Your tour guide has provided a focused reading approach to these topics.
Regression Analysis by Example Chapter 3,pages 57 – 81

Note p61 estimate for sigma squared = SSE/(n-p-1), p 67 BLUE, p 68 r squared and adjusted r squared, p 69 null hypothesis beta = 0, p71 FM, RM and nested models, p72 F ratio, p73 null hypothesis that all beta = 0, p75 principle of parsimony p 77 Remark #2.

Python Data Science Handbook, Chapter 2 pages 33-96 and Chapter 5 pages 390-405

Note, for Chapter 2, there are some great tools for data sorting, concatenation and more. Use this chapter as a reference tool as needed.
Same for Chapter 5, mainly try the code on p 391 for linear regression.

Discussion Board

Join in the week-specific discussion board forum. Your participation in both posting and responding will help you gain a better understanding of the material and provide interaction with other students.

Quiz #1 & Quiz#2

Assignments – Assignment #1(300 points)

Assignment #1: Simple and Multiple Regression Model Building is due Sunday.

Week 3: – Topic: Model Validation

Learning Objectives

After this week, 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, how to detect outliers, and how outliers can affect the regression fit.
- Define the statistical concept of leverage, how leverage is computed, how to use leverage estimates to detect outliers, and 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

Your tour guide has provided a focused reading approach to these topics.
Regression Analysis by Example Chapter 4,pages 93 – 123
Note Chapter 4 p94-96 linear regression assumptions for the model, errors, predictor variables and collinearity, p 96-97 residuals, p99 use of plots and influential observations, p105 checking assumptions with plots, p106 leverage, influence and outliers, p109 masking and swamping, and p111 Cook's distance.

Python Data Science Handbook, Chapter 3 pages 97-126
Note, Pandas package, p102 dataframes, and p119-126 missing data, and NaN.

Discussion Board

Join in the week-specific discussion board forum. Your participation in both posting and responding will help you gain a better understanding of the material and provide interaction with other students.

Assignments – No Assignment Due
No Sync Session

Week 4 – Topic: Variable Transformations

Learning Objectives

After this week, 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.
- 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

Your tour guide has provided a focused reading approach to these topics.
Regression Analysis by Example Chapter 5, pages129 – 150
Note p129 indicator or dummy variables, and p133 interaction variables.
Regression Analysis by Example Chapter 6, pages 163 – 186
Note p165 transformations.

If you need it, Python Data Science Handbook, Chapter 3 pages 128-158

Discussion Board

Join in the week-specific discussion board forum. Your participation in both posting and responding will help you gain a better understanding of the material and provide interaction with other students.

Assignment – No Assignment Due

No Sync Session

Week5 – Topic: Automated Variable Selection

Learning Objectives

After this week, the student will be able to:

- 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

Your tour guide has provided a focused reading approach to these topics.
Regression Analysis by Example Chapter 11,pages 299 – 328
Note p 304 Mallows Cp, p305 AIC and BIC, p307-309 forward selection, backward elimination and stepwise.

Use the Python Handbook as reference tool if needed.
Python Data Science Handbook, Chapter 3 pages 158-215

Discussion Board

Join in the week-specific discussion board forum. Your participation in both posting and responding will help you gain a better understanding of the material and provide interaction with other students.

Quiz #3 #4 & #5

Assignments – Assignment #2 (300 Points)

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

No sync session scheduled

Week6 – Topic: Multicollinearity and Principal Components Analysis

Learning Objectives

After this week, 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

Your tour guide has provided a focused reading approach to these topics.
Regression Analysis by Example Chapter 9, pages 233 – 251
Regression Analysis by Example Chapter 10, pages 259 – 287

For the assigned reading, focus on the topics mentioned here:
Chapter 9 P233 orthogonal predictor variables, p238 discussion re predictor variables with low t-values, p235 residual plot, 239 pairwise scatter plot and correlations, p245 signs of collinearity, and p248-250 VIF > 10.
Chapter 10 p259 predictor variables transformed to a set of orthogonal variables.

Some other reading in case you would like more. It is not required that you read these but the more examples you see the more places you will find to use your skills. (I will post these for you):
PCA Chapter 3 by Everett & Dunn
Factor Analysis Chapter 12 by Everett & Dunn
Factor Analysis by Stoetzel
Factor Analysis by Morrison Chapter 7
Eigenvalues and Eigenvectors
PCA Examples

Optional (if you need it)
Python Data Science Handbook, Chapter 5 pages 433-445

Discussion Board
Join in the week-specific discussion board forum. Your participation in both posting and responding will help you gain a better understanding of the material and provide interaction with other students. See Ratner article on variable selection.

Assignments – No assignment due.

Week7 – Topic: Exploratory Factor Analysis

Learning Objectives

After this week, the student will be able to:

- 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 Python.
- Fit, interpret, and validate a factor analysis.
- Apply factor rotations to increase factor interpretation.
- Use the output from a factor estimation 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

Readings were posted for week 6.

Discussion Board

Join in the week-specific discussion board forum. Your participation in both posting and responding will help you gain a better understanding of the material and provide interaction with other students.

Quiz #6

Assignments – Assignment #3 (100 points)

PCA and Factor Analysis is due Sunday at 11:55 p.m. (Central Time).

No Sync Session Scheduled

Week8 – Topic: Cluster Analysis

Learning Objectives

After this week, 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.

Optional Reading

Python Data Science Handbook, Chapter 5 pages 462-476

Discussion Board

Join in the week-specific discussion board forum. Your participation in both posting and responding will help you gain a better understanding of the material and provide interaction with other students.

Assignments – No assignment due.

Sync Session- None

Week9 – Topic: Multivariate Data Analysis

Learning Objectives

After this week, 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

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

Discussion Board - None

Assignments – Assignment #4 (100 points) & Assignment 4A (100 points)

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

No Sync Session