Engineering Faculty Document No. 58-16
April 18, 2018

TO: The Faculty of the College of Engineering

FROM: The Faculty of the Weldon School of Biomedical Engineering

RE: New Graduate Course BME 50100 Multivariate Analyses in Biostatistics

The faculty of the Weldon School of Biomedical Engineering has approved the following new course. This action is now submitted to the Engineering Faculty with a recommendation for approval.

BME 50100 Multivariate Analyses in Biostatistics
Term Offered: Fall or Spring, Lecture 3, Cr. 3.
Restriction: Graduate Level OR STAT 35000 OR STAT 51100 OR IE 330

Description: This course focuses on fundamental principles of multivariate statistical analyses in biostatistics, including multiple linear regression, multiple logistic regression, analysis of variance, and basic epidemiology. The fundamental theories are applied to analyze various biomedical applications ranging from laboratory data to large-scale epidemiological data.

Reason: This course serves as the focal point of discourse for students in biomedical engineering and other biomedical science disciplines requiring an in-depth understanding of statistical analyses and designs of experimental studies. To successfully validate novel technologies, to clearly understand published results, and to effectively communicate with medical communities, biomedical researchers are required to have a solid understanding of study designs and rigorous statistical methods. This course has been taught eight times with an average of 28 students per offering as one of the required quantitative courses in BME.

George R. Wodicka

George R. Wodicka,
Dane A. Miller Head and Professor
Weldon School of Biomedical Engineering

Detailed Graduate Course Proposal for Academic Review

Note: The detailed course proposal is intended for academic review by the appropriate area committee of the Graduate Council. It supplements the Form 40G that is intended for administrative review of the Graduate School and Registrar.

To: Purdue University Graduate Council

From: Faculty Member: Young Kim
Department: Weldon School of Biomedical Engineering
Campus: West Lafayette

Date: April 19, 2018

Subject: Proposal for New Graduate Course

Contact for information if questions arise:
Name: Young Kim
Phone: 6-2445
Email: youngkim@purdue.edu
Address: MJIS 3027

Course Number: BME 50100
Course Title: Multivariate Analyses in Biostatistics
Short Title: Biostatistics

Course Description:

This course focuses on fundamental principles of multivariate statistical analyses in biostatistics, including multiple linear regression, multiple logistic regression, analysis of variance, and basic epidemiology. The fundamental theories are applied to analyze various biomedical applications ranging from laboratory data to large-scale epidemiological data.

A. Justification for the Course

Justification of the need for the course

In the early days of biomedical sciences and technologies, statistical analyses and designs of studies were not critical as most studies were aimed to demonstrate their feasibilities. To successfully validate novel technologies and concepts to practice and to effectively communicate with medical communities, 21st century biomedical engineers are now required to look into well-designed studies and rigorous statistical analyses. Because several variables are included in typical biomedical studies, multivariate analyses play a key role. In addition, to effectively cover the concepts of multivariate statistical analyses, data sets and issues from various sources include:

- HIV screening from Centers for Disease Control and Prevention (CDC)
- Framingham Heart Study (http://www.framinghamheartstudy.org)
- Breast cancer screening program in the health insurance plan (HIP) of New York
- CDC recommendation for HIV tests from Morbidity and Mortality Weekly Report

Such case studies and examples from epidemiologic and biomedical studies are rigorously analyzed using a statistical software package (e.g. STATA, SAS, SPSS, and R).

The most relevant course offered by other departments is STAT 512 - Applied Regression Analysis at the Department of Statistics. The course description is "Thorough applied course in regression and analysis of variance including experience with the SAS statistical software package. Topics include inference in simple and multiple linear regression, residual analysis, transformations, polynomial regression, model building with real data, nonlinear regression, one-way and two-way analysis of variance, multiple comparisons, fixed and random factors, analysis of covariance." Thus, the only overlap between this proposed course and STAT 512 is linear regression, which accounts for approximately one week lecture. In addition, examples and questions in STAT 512 are not designed to be specific to cover common issues and problems in biomedical research.

Justification that course will be taught at a graduate level

Multivariate regression analyses combined with basic epidemiology can be highly beneficial for enhancing the awareness of critical issues in translational research and technology transfer in graduate students in BME. The course is designed for all graduate students in BME to build a foundation of multivariate statistical analyses, to analyze large data with multiple variables, to formulate appropriate models from biostatistical perspectives, and to correctly interpret statistical estimates. Thus, a 500-level course would be ideal to cover the fundamental theories and their applications to various biomedical problems ranging from laboratory data to large-scale epidemiologic data. This course will also allow the students to understand how major types of bias and obstacles in designing and conducting clinical trials.

Justification of the demand for the course

- Anticipated enrollment
  - Undergraduate 10
  - Graduate 30

Justification for online delivery

This course was already delivered online in Fall 2000, Fall 2011, Fall 2012, Fall 2015, and Fall 2016.

B. Learning Outcomes and Methods of Assessment

Learning Outcomes	Assessment Methods
understand fundamental concepts in multivariate regression analyses	quizzes, midterm, and final exams
translate research questions into suitable multivariate frameworks	quizzes, midterm, and final exams
design studies with appropriate statistical power	quizzes, midterm, and final exams
compute and interpret appropriate statistical estimates from multivariate analyses	quizzes, midterm, and final exams
use a major statistical software package (e.g. STATA, SAS, SPSS, and R) to perform large-size data analyses	quizzes, midterm, and final exams

- These exams will be designed to assess students' knowledge of multivariate biostatistical analyses, their skills in calculating and interpreting statistical computations, and their knowledge gained from the supplementary course readings.

Final Grading Criteria

Assessment Methods (should match method types in the previous table)	Weight Toward Final Course Grade
Class participation, assignment, and quizzes	30%
Midterm exam	40%
Final exam	40%

Methods of Instruction

Class Hrs/Week	Method of Instruction	Contribution to Outcomes
	Lecture	Lectures will be used to help the student to achieve the learning outcomes. These include hands-on analyses of basic principles of concepts covered during lectures.

C. Prerequisite(s)

- Graduate Student status OR STAT 35000 OR STAT 51100 OR IE 330

D. Course Instructor(s)

Name	Rank	School, dept., or center	Graduate Faculty or expected date
Young Kim	Associate Professor	BME	Yes
[click here and type name]	[Rank]	[Dept]	[Yes, no, or date]
[click here and type name]	[Rank]	[Dept]	[Yes, no, or date]
[click here and type name]	[Rank]	[Dept]	[Yes, no, or date]
[click here and type name]	[Rank]	[Dept]	[Yes, no, or date]

Young Kim has experience and training in epidemiology and biostatistics as a bioengineer. In particular, in addition to his PhD in biomedical engineering, he has an MS degree in clinical investigation during his postdoctoral fellow training supported by NCI and has several papers in the field of biostatistics.

E. Course Schedule or Outline

Week 1-2

Introduction to course
- Statistical software packages
- Course formats and evaluations

Review of inferential statistics

Introduction to basic epidemiological concepts
- Epidemiology in engineering
- Basic study designs in analytical epidemiology
- Defining disease burden
- Comparing groups (e.g. relative risk and odd ratio)
- Adjustment of rates
- Confounding

Week 3-4

Analysis of variance (ANOVA)
- Within group variance
- Between group variance
- Hypothesis testing

Simple linear regression I
- Basic principle for simple linear regression
- Correlation coefficient
- Significance test
- Goodness of fit
- Hypothesis test and confidence intervals of parameters
- Assumption in simple linear regression

Simple linear regression II
- Interpretation of simple linear regression
- Regression line fitted value, residual, and influence
- Confounder and interaction in simple linear regression
- Transforming variables
- Overview of "Framingham Heart Study"

Week 5- 7

Correlation analysis
- Basic principle of correlation analysis
- Relationship between linear regression and correlation analysis
- Inference of correlation coefficient
- Partial correlation coefficient

Multiple linear regression I
- Basic principle for multiple linear regression
- R^2 statistics
- Confounding and interaction in multiple linear regression
- Correlation in multiple linear regression
- Defining dummy variables

Multiple linear regression II
- Hypothesis testing
- Assumption in multiple linear regression
- Residual analysis
- Confidence interval for prediction and forecast

- Predictor selection
- Interpretation of regression model output
- Translation of research question into regression models
- Relationship between ANOVA and multiple linear regression

Week 8

Screening and diagnostic tests
- Principles and practice of screening
- Bias in screening
- Screening test performance and validity matrices
- Effect of multiple tests

Week 9-12

Inferential statistics for proportion
Simple logistic regression I
- Basic principle of simple logistic regression
- Relationship between logistic regression and linear regression
- Logit transform and binomial distribution
- Probability and odds
- Assumption in logistic regression

Simple logistic regression II
- Maximum likelihood estimation
- Hypothesis testing
- Confidence interval
- Confounding and interaction in simple logistic regression
- Contingency tables

Multiple logistic regression I
- Mantel-Haenszel estimate
- Basic principle of multiple logistic regression
- Linear vs logistic regression
- Inference in multiple logistic regression
- Test of interaction and confounding
- Defining dummy variables

Multiple logistic regression II
- Logistic regression in case-control study and cohort study
- Logistic regression in screening study
- Receiver operating characteristic (ROC) curve
- Goodness of fit test
- Assumptions in logistic regression
- Residual analysis
- Translation of research question into regression model
- Interpretation of regression model output

Week 13-14

Analysis of variance
- Basic principles of analysis of variance (ANOVA)
- Relationship between (ANOVA) and linear regression

Multiple comparisons for ANOVA
-Bonferroni Method
-Tukey method
-Scheffe method

- Non-parametric ANOVA

Week 15

Introduction to survival analysis
- Introduction and basic principles
- Kaplan-Meier curve
- Logrank test
- Survival analysis vs other regression analyses

Week 16

Issues on statistical power and sample size estimation
- Type I and type II errors
- Statistical power
- Sample size calculation
- Designing studies

F. Reading List (including course text)

Primary Reading List

- Statistical Modeling for Biomedical Researchers. Cambridge University Press, 2nd edition, 2009. ISBN10: 0521614805, ISBN-13: 978-0521614801

Secondary Reading List

- Vittinghoff E, Glidden DV, Shiboski SC, McCulloch CE. Regression Methods in Biostatistics. Springer; 2nd edition, 2012.
- Kachigan SK. Statistical Analysis: an Interdisciplinary Introduction to Univariate & Multivariate Methods, Radius Press; 2nd edition, 1991.

G. Library Resources

Name of journal, proceedings, book, video, or other acquisition	Already in Libraries?
Purdue libraries have the books	Yes

H. Course Syllabus (now required)

BME 595 Multivariate Analysis in Biostatistics (Biostatistics)
3.0 Credit

Time: Fall 2017, T/TH 9:00 am – 10:15 am

Location: MJIS 1083

Office hours: After lectures

Course Instructor
Young Kim, PhD, MSCI
Associate Professor
Weldon School of Biomedical Engineering
Tel: 765-496-2445
Email: youngkim@purdue.edu

Prerequisite: IE 330 Probability and Statistics in Engineering II or equivalent basic statistics course.

Description
This course focuses on fundamental principles of multivariate statistical analyses in biostatistics, including multiple linear regression, multiple logistic regression, analysis of variance, and basic epidemiology concepts. The fundamental theories are applied to analyze various biomedical applications ranging from laboratory data to large-scale epidemiological data. In particular, this course focuses on multivariate statistical analyses, which involve more than one variable and take into account several variables on the responses of interest.

Motivation
In the early days of biomedical sciences and technologies, statistical analyses and designs of studies were not critical as most studies were aimed to demonstrate their feasibilities. To successfully validate novel technologies to practice and to effectively communicate with medical communities, 21st century biomedical researchers and engineers are now required to look into well-designed studies and rigorous statistical analyses. Multivariate regression analyses combined with basic epidemiology can be highly beneficial in translational research and technology transfer.

Approach
This course is an intermediate biostatistics course, based on mathematical formations, explanations of key concepts, and hands-on analyses (neither a cookbook approach nor a sophisticated mathematical approach).

Case and example
To provide solid understandings of multivariate biostatistical analyses, current biomedical research problems with large sample sizes are also rigorously analyzed using a statistical software package. Case studies and examples are based on real data sets and real issues from various sources including the followings:

1. HIV screening from Centers for Disease Control and Prevention (CDC)

2. Age standardization of death rates: implementation of the year 2000 standard, National vital statistics reports
3. Framingham Heart Study (http://www.framinghamheartstudy.org)
4. Breast cancer screening program in the health insurance plan (HIP) of New York
5. CDC recommendation for HIV tests from Morbidity and Mortality Weekly Report
6. Spectroscopic detection of colon cancer

Objective

The course is designed for graduate students to build a foundation of multivariate statistical analyses, to analyze large data with multiple variables, to formulate appropriate models from different perspectives, and to correctly interpret statistical estimates. By the completion of this course, students will be able to:

1) **understand fundamental concepts in multivariate regression analyses,**
2) **apply data analyses using an advanced statistical software package,**
3) **formulate research questions into suitable multivariate frameworks, and**
4) **analyze appropriate statistical estimates from multivariate analyses.**

Lecture Topics and Schedule

Week 1

Introduction to course
- Statistical software packages
- Course formats and evaluations

Review of inferential statistics
- Terminology
- Point estimation and interval estimation
- Central limit theorem
- Hypothesis testing

Example of population variance and sample variance

Week 2

Introduction to basic epidemiological concepts
- Epidemiology in engineering
- Basic study designs in analytical epidemiology
- Defining disease burden
- Comparing groups (e.g. relative risk and odd ratio)
- Adjustment of rates
- Confounding

AIDS example from "The band played on" by Randy Shilts

Example of age adjustment from "Age standardization of death rates: implementation of the year 2000 standard, National vital statistics reports"

Discussion in class
- Each student will present a study design of his/her own interest and find a corresponding article

Week 3

Analysis of variance (ANOVA)
- Within group variance
- Between group variance
- Hypothesis testing

Simple linear regression I
- Basic principle for simple linear regression
- Correlation coefficient
- Significance test
- Goodness of fit
- Hypothesis test and confidence intervals of parameters
- Assumption in simple linear regression

<u>Hypothetical example of systolic blood pressure and age</u>

Week 4

Simple linear regression II
- Interpretation of simple linear regression
- Regression line fitted value, residual, and influence
- Confounder and interaction in simple linear regression
- Transforming variables
- Overview of "Framingham Heart Study"

<u>Example from "Framingham Heart Study"</u>

Week 5

Correlation analysis
- Basic principle of correlation analysis
- Relationship between linear regression and correlation analysis
- Inference of correlation coefficient
- Partial correlation coefficient

Multiple linear regression I
- Basic principle for multiple linear regression
- R^2 statistics
- Confounding and interaction in multiple linear regression
- Correlation in multiple linear regression
- Defining dummy variables

<u>Example from "Framingham Heart Study"</u>

Week 6-7

Multiple linear regression II
- Hypothesis testing
- Assumption in multiple linear regression
- Residual analysis
- Confidence interval for prediction and forecast
- Predictor selection
- Interpretation of regression model output
- Translation of research question into regression models
- Relationship between ANOVA and multiple linear regression

<u>Example from "Framingham Heart Study"</u>

Midterm Exam consisting of large data sets from current biomedical studies, hands-on calculations, and critical literature analysis from a published article

Week 8

Screening and diagnostic tests
- Principles and practice of screening
- Bias in screening
- Screening test performance and validity matrices

- Effect of multiple tests

Example from "Breast cancer screening program in the health insurance plan (HIP) of New York"

Example from "CDC recommendation for HIV tests from Morbidity and Mortality Weekly Report"

Discussion in class

- Exercise on HIV screening "Screening for Antibody to the Human Immunodeficiency Virus" from Centers for Disease Control and Prevention

Week 9

Inferential statistics for proportion

Simple logistic regression I

- Basic principle of simple logistic regression
- Relationship between logistic regression and linear regression
- Logit transform and binomial distribution
- Probability and odds
- Assumption in logistic regression

Example of age and coronary heart disease

Week 10

Simple logistic regression II

- Maximum likelihood estimation
- Hypothesis testing
- Confidence interval
- Confounding and interaction in simple logistic regression
- Contingency tables

Example of age and coronary heart disease

Week 11

Multiple logistic regression I

- Mantel-Haenszel estimate
- Basic principle of multiple logistic regression
- Linear vs logistic regression
- Inference in multiple logistic regression
- Test of interaction and confounding
- Defining dummy variables

Example of lung cancer and esophageal cancer

Week 12

Multiple logistic regression II

- Logistic regression in case-control study and cohort study
- Logistic regression in screening study
- Receiver operating characteristic (ROC) curve
- Goodness of fit test
- Assumptions in logistic regression
- Residual analysis
- Translation of research question into regression model
- Interpretation of regression model output

Example from "Primary biliary cirrhosis trial"

Example from "Spectroscopic detection of colon cancer"

Week 13

Analysis of variance
- Basic principles of analysis of variance (ANOVA)
- Relationship between (ANOVA) and linear regression

Multiple comparisons for ANOVA
-Bonferroni Method
-Tukey method
-Scheffe method
- Non-parametric ANOVA

<u>Example from "Effect of estrogen receptor genotype at diagnosis"</u>

Week 14

Checking assumptions in ANOVA
-Normality
-Homoscedasticity

Non-parametric Analysis of Variance
-Wilcoxon Rank-sum test
-Kruskal-Wallis test
-Multiple comparisons

Week 15

Repeated measure analysis of variance
-Two-stage analysis

<u>Example from "Photocarcinogenesis animal study"</u>

Survival analysis
- Introduction and basic principles
- Kaplan-Meier curve
- Logrank test
- Survival analysis vs other regression analyses

<u>Hypothetical example of treatment and survival time</u>

Week 16

Final review

Final Exam consisting of large data sets from current biomedical studies, hands-on calculations, and critical literature analysis from a published article

<u>Grade Assessment:</u>

33%	-	Class participation and weekly assignment
33%	-	Midterm exam
34%	-	Final exam

<u>Grading:</u>

Total $\geq$ 90%:	A
80% $\leq$ Total < 90%:	B
70% $\leq$ Total < 80%:	C
60% $\leq$ Total < 70%:	D

<u>Homework assignment</u>

For this course, students will be evaluated in terms of their performance on weekly assignment. The assignment will be evaluated critically. The assignment will include hands-on analyses of basic principles of concepts covered during lectures using small sample examples.

Statistical software package
This course will also conduct basic operations in large-scale statistical analyses with statistical software packages (e.g. STATA). Students will have access to the packages in BME computer lab. Six-month ($45) STATA license is also available to purchase his/her own student version (http://www.stata.com/order/new/edu/gradplans/student-pricing). STATA is chosen for the breadth and depth of its statistical methods, its ease of use, and its good documentation. However, students can use other software packages such as SAS, SPSS, and R (free statistical software). Overall, minimal help on STATA will be provided, given that this course is intended to teach biostatistics not a package. It should be noted that STATA is a user-friendly easy-to-use statistical software package that is commonly used in clinical and biomedical research.

Examinations
There will be one take-home midterm and one take-home final exam scheduled during this course. These tests will be designed to assess students' knowledge of multivariate biostatistical analyses, their skills in calculating and interpreting statistical computations, and their knowledge gained from the supplementary course readings. Each exam will cover different content; hence, the final exam will not be cumulative. However, the format of examinations is subject to change, depending on students' performance and participation.

Textbook:

Dupont WD. Statistical Modeling for Biomedical Researchers. Cambridge University Press, 2nd edition, 2009. http://biostat.mc.vanderbilt.edu/dupontwd/wddtext/index.html

Other Useful References:

Vittinghoff E, Glidden DV, Shiboski SC, McCulloch CE. Regression Methods in Biostatistics. Springer; 2nd edition, 2012.
Kachigan SK. Statistical Analysis: an Interdisciplinary Introduction to Univariate & Multivariate Methods. Radius Press; 2nd edition, 1991.

UCLA Statistical Consulting Group: http://www.ats.ucla.edu/stat

Course Description for Multivariate analyses in biostatistics

This course focuses on fundamental principles of multivariate statistical analyses in biostatistics, including multiple linear regression, multiple logistic regression, analysis of variance, and basic epidemiology. The fundamental theories are applied to analyze various biomedical applications ranging from laboratory data to large-scale epidemiological data.

Learning Outcomes for Multivariate analyses in biostatistics

- understand fundamental concepts in multivariate regression analyses
- translate research questions into suitable multivariate frameworks
- design studies with appropriate statistical power
- compute and interpret appropriate statistical estimates from multivariate analyses
- use a major statistical software package (e.g. STATA, SAS, SPSS, and R) to perform large-size data analyses