南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

课程详述
COURSE SPECIFICATION

以下课程信息可能根据实际授课需要或在课程检讨之后产生变动。如对课程有任何疑问，请联系授课教师。

The course information as follows may be subject to change, either during the session because of unforeseen circumstances, or following review of the course at the end of the session. Queries about the course should be directed to the course instructor.

1.	课程名称 Course Title	大数据统计与计量分析方法 Big Data Statistics and Econometric Methods
2.	授课院系 Originating Department	信息系统与管理工程系 Division of Information Systems & Management Engineering
3.	课程编号 Course Code	MIS 210
4.	课程学分 Credit Value	3
5.	课程类别 Course Type	专业选修课 Major Elective Courses
6.	授课学期 Semester	春季 Spring
7.	授课语言 Teaching Language	英文 English
8.	授课教师、所属学系、联系方式（如属团队授课，请列明其他授课教师） Instructor(s), Affiliation& Contact （For team teaching, please list all instructors）	Moris Strub Department of Information Systems and Management Engineering College of Business Taizhou Building Rm 501-5(D) strub@sustech.edu.cn
9.	实验员/助教、所属学系、联系方式 Tutor/TA(s), Contact	待公布 To be announced
10.	选课人数限额(可不填) Maximum Enrolment （Optional）	

11. 授课方式 Delivery Method	讲授 Lectures	习题/辅导/讨论 Tutorials	实验/实习 Lab/Practical	其它(请具体注明) Other（Please specify）	总学时 Total
学时数	32		32		64

Credit Hours					

12.	先修课程、其它学习要求 **Pre-requisites or Other Academic Requirements**	**MA212 概率论与数理统计 Probability and Statistics**
13.	后续课程、其它学习规划 **Courses for which this course is a pre-requisite**	
14.	其它要求修读本课程的学系 **Cross-listing Dept.**	

教学大纲及教学日历 SYLLABUS

15. **教学目标 Course Objectives**

This is a major elective course where students will become familiar with important statistical tools and econometric methods required to analyse big data in a modern business environment. This course aims to help students to gain the skills to operate as a data scientist at a sophisticated data-driven firm. We will introduce the basic concepts, theories, and methods of Econometrics and Big Data Statistics such as statistical analysis, econometric analysis, data modeling, model selection, and optimal decision making. Besides learning about the underlying theoretical models and tools, a lot of emphasis will be placed on real world applications and economic interpretations. Students will learn how to apply R to perform big data analysis on real-world economic data.

16. **预达学习成果 Learning Outcomes**

After finishing this course, students should be able to

a) know about the methodologies of econometrics and the foundations and techniques of big data statistics;

b) be familiar with common business problems and able to perform data analytics to drive better business decision-making;

c) be familiar with the mathematical and economic theories, models, methods and algorithms for solving big data problems relevant for the modern business world;

17. **课程内容及教学日历 （如授课语言以英文为主，则课程内容介绍可以用英文；如团队教学或模块教学，教学日历须注明主讲人）**
Course Contents (in Parts/Chapters/Sections/Weeks. Please notify name of instructor for course section(s), if this is a team teaching or module course.)

Lecture (32 hours)

Lec1 – Economic Data, Big Data, and the Nature of Econometrics [BDS0,ISL1,IE1]

In this lecture, students will be introduced Big Data and its characteristics, and discuss the scope of econometrics and raises general issues that arise in the application of econometric methods.

Lec2 – Uncertainty and Probability [BDS1,IE-AppendixB&C]

In this lecture, students will learn the concept of uncertainty and review basic results from probability theory and

statistics

Lec3 – Linear Regression: Estimation I [BDS2,ISL3,IE2&3]

In this lecture, students will learn what is linear model, how to interpret the simple regression model and know about that the simple regression model has limitations as a general tool for empirical analysis.

Lec4 – Linear Regression: Estimation II [BDS2,ISL3,IE2&3]

In this lecture, students will learn the multiple regression models and further discuss the advantages of multiple regressions over simple regressions. Students will also know about how to estimate the parameters in the multiple regression models using the method of ordinary least squares and describe various statistical properties of the OLS estimators.

Lec5 – Linear Regression: Inference [ISL3,IE4]

In this lecture, students will continues learning multiple regression analysis, and turn to the problem of testing hypotheses about the parameters in the population regression model, which includes that testing about individual parameters, how to test a single hypothesis involving more than one parameter, and test multiple restrictions.

Lec6 – Linear Regression: Big Data Asymptotics [IE5]

In this lecture, students will learn the asymptotic properties or large sample properties of estimators and test statistics, and know that even without the normality assumption, t and F statistics have approximately t and F distributions, at least in large sample sizes.

Lec7 – Linear Regression: Qualitative Information [IE7]

In this lecture, students will learn to discuss qualitative independent variables, and know about how qualitative explanatory variables can be easily incorporated into multiple regression models. Students will also learn to discuss a binary dependent variable.

Lec8 – Linear Regression: Heteroskedasticity [IE8]

In this lecture, students will review the consequences of heteroskedasticity for ordinary least squares estimation, and learn the available remedies when heteroskedasticity occurs, and also know about how to test for its presence.

Lec9 – Resampling Methods [BDS1,ISL5]

In this lecture, students will know the resampling methods: cross-validation and bootstraps, and be introduced K-fold cross-validation, nonparametric and parametric bootstraps.

Lec11 – Model Selection [BDS3,ISL6]

In this lecture, students will know why we need fitting procedures other than least squares. Best subset selection, forward selection and backward selection will be introduced. Students will also learn two common approaches that used to select the best model, including indirect estimation of test error, Cp, AIC, BIC, adjusted R^2, and indirect estimation of test error, validation and cross-validation to select the best model.

Lec12 – Regularization [BDS3,ISL6]

In this lecture, students will know why we need regularization to avoid overfitting in analysis. Students will learn regularization paths. Ridge regression and lasso regression paths will be introduced. How to use cross-validation to select the best model will be reviewed.

Lec13 – Classification I [BDS4,ISL4]

In this lecture, students will study the approaches for predicting qualitative responses - classification. And I will explain why linear regression is not suitable in the case of a qualitative response. Students will be introduced K nearest neighbours, discuss the relationship between probabilities and classification, and review logistic regression.

Lec14 – Classification II [BDS4,ISL4]

In this lecture, students will study multinomial logistic regression, distributed multinomial regression and MapReduce

framework algorithm.

Lec15 – Clustering [BDS7,ISL10]

In this lecture, students will know the basic concepts of clustering and differences between clustering and classification. And students will study the concepts of K-means clustering and hierarchical clustering and their algorithms. Besides, practical issues in clustering will be introduced.

Lec16 – Factorization [BDS7,ISL10]

In this lecture, students will learn the concepts of principle components, and the procedure of principle component analysis. Also, student will be introduced principle components regression (PCR) and its algorithm, and why PCR is fruitful.

Lab (32 hours)

Lab1- Introduction to "R"- I [ISL2]

In this lab, "R", its advantages and resources will be introduced. Students will know how to install and use R. And the basic commands and graphics functions will be learned.

Lab2- Introduction to "R"- II [ISL2]

In this lab, students will know how to examine part of a set of data and how to load data.

Lab3- Linear Regression: Estimation I [ISL3]

In this lab, students will be Introduced library and know how to conduct simple linear regression with R

Lab4- Linear Regression: Estimation II [ISL3]

In this lab, students will know how to conduct multiple linear regression using lm( ) function with R and the usage of the summary( ) function.

Lab5- Linear Regression: Inference [ISL3,IE4]

In this lab, students will learn how to perform statistical tests about parameters of a multiple linear regression model with R.

Lab6- Linear Regression: Big Data Asymptotics [IE5]

In this lab, students will know how to conduct t-test and F-test when facing a large size of sample with R.

Lab7- Linear Regression: Qualitative Information [IE7]

In this lab, students will learn paths of qualitative predictor with R.

Lab8- Linear Regression: Heteroskedasticity [IE8]

In this lab, students will know how to perform test of heteroskedasticity for ordinary least squares estimation with R.

Lab9- Resampling Methods [BDS1,ISL5]

In this lab, students will know how to conduct Leave-one-out cross validation, k-fold cross-validation with R. Also, students will learn how to perform bootstrap to estimate the accuracy of a statistic of interest and the accuracy of a linear regression model with R.

Lab11- Model Selection [BDS3,ISL6]

In this lab, students will know how to conduct best subset selection with R and the associated function. Moreover, students will learn to perform forward or backward stepwise selection using regsubsets( ) function with R.

Lab12- Regularization [BDS3,ISL6]

In this lab, students will know how to perform ridge regression and lasso regression using glmnet( ) function with R.

Lab13- Classification I [BDS4,ISL4]

In this lab, students will learn to perform K-nearest neighbours using knn( ) function with R.

Lab14- Classification II [BDS4,ISL4]

In this lab, students will learn to conduct logistic regression glm( ) function with R.

Lab15- Clustering [BDS7,ISL10]

In this lab, students will learn how to conduct K-means clustering and hierarchical clustering with R.

Lab16- Factorization [BDS7,ISL10]

In this lab, principle component analysis will be introduced with R. NCI60 Data will be used to conduct PCA and hierarchical clustering analysis.

18. 教材及其它参考资料 Textbook and Supplementary Readings

Matt Taddy, Business Data Science, McGraw Hill, 2019. [BDS]

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, 2013. [ISL]

Jeffrey M. Wooldridge, Introductory Econometrics: A Modern Approach, 6th edition, 清华大学出版社（影印版, 2017. [IE]

课程评估 ASSESSMENT

19.

评估形式 Type of Assessment	评估时间 Time	占考试总成绩百分比 % of final score	违纪处罚 Penalty	备注 Notes
出勤 Attendance		10		
课堂表现 Class Performance				
小测验 Quiz				
课程项目 Projects				
平时作业 Assignments		30		
期中考试 Mid-Term Test		25		
期末考试 Final Exam		35		
期末报告 Final Presentation				
其它（可根据需要改写以上评估方式） Others (The above may be modified as necessary)				

20. 记分方式 **GRADING SYSTEM**

☑ A. 十三级等级制 **Letter Grading** ☐ B. 二级记分制（通过/不通过） **Pass/Fail Grading**

课程审批 REVIEW AND APPROVAL

21. **本课程设置已经过以下责任人/委员会审议通过**
This Course has been approved by the following person or committee of authority
