SMU Classification: Restricted

GLOBAL SUMMER PROGRAMME 2026

DS601 STATISTICAL LEARNING

Instructor Name : Dr. Shew Fan Liu
Title : Senior Lecturer of Statistics
Email : sfliu@smu.edu.sg
Office : School of Economics, 5044

COURSE DESCRIPTION

This course provides students with a rigorous and comprehensive foundation in modern, data-driven statistical modelling, bridging classical statistical theory with contemporary machine learning methods and emphasising the principles that govern predictive modelling, inference, and reliable performance on new, unseen data. Programming is integrated throughout the course.

Designed for students in data science, applied economics, and related quantitative disciplines, the course aims to develop statistical literacy, computational proficiency, and the critical evaluation skills needed to analyse rich, real-world datasets and construct reliable predictive models.

No prior programming experience is required.

LEARNING OBJECTIVES

By the end of this course, students will be able to:

1. Describe and explain the fundamental principles of statistical modelling, inference, and prediction in both classical and modern data-driven frameworks.
2. Distinguish between explanatory and predictive models, and justify when each is appropriate in applied analysis.
3. Apply and evaluate statistical and machine learning methods in Python to analyse rich, real-world datasets and assess model performance using cross-validation, regularisation, and related techniques.
4. Compare and justify modelling choices based on theoretical considerations, data characteristics, and model generalisability.
5. Synthesise and communicate results clearly and accurately to both technical and non-technical audiences.
6. Demonstrate sound computational and ethical practices, including code readability, reproducibility, and responsible data use in statistical modelling workflows.

PRE-REQUISITES

Introductory Statistics

SMU Classification: Restricted

ASSESSMENT METHODS

Class Participation	10%
Group Assignments	30%
Individual Assignments	40%
Exam	20%
Total	**100%**

INSTRUCTIONAL METHODS AND EXPECTATIONS

The course will be conducted on campus, and students are expected to attend all scheduled sessions. If unforeseen circumstances prevent attendance, they should notify the instructor promptly and provide a valid reason.

The course will be delivered through a combination of lectures and lab sessions. Lectures will focus on introducing key concepts, theoretical foundations, and illustrative examples, while lab sessions will emphasize practical implementation, data analysis, and interpretation of results. Python will serve as the primary programming language, with optional exposure to R for comparative or supplementary purposes.

Class Participation:
Active participation in discussions during classes is highly encouraged.

Assignments:
There will be a series of assignments over the duration. Individual assignments will be attempted in class after each lab session. There will be two group assignments that will be presented in class.

Final Examination:
The final examination will be a 2-hour computer-based exam, conducted in an open-book format.

CONSULTATIONS

By appointment

RECOMMENDED TEXT AND READINGS

Textbooks:

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J., (2023). *An Introduction to Statistical Learning with Applications in Python* (1st ed.). Springer.

Readings:

Amazon Machine Learning University, available for **free** at: https://mlu-explain.github.io

SMU Classification: Restricted

UNIVERSITY POLICIES

Academic Integrity

All acts of academic dishonesty (including, but not limited to, plagiarism, cheating, fabrication, facilitation of acts of academic dishonesty by others, unauthorized possession of exam questions, or tampering with the academic work of other students) are serious offences.

All work (whether oral or written) submitted for purposes of assessment must be the student's own work. Penalties for violation of the policy range from zero marks for the component assessment to expulsion, depending on the nature of the offense.

When in doubt, students should consult the instructors of the course. Details on the SMU Code of Academic Integrity may be accessed at https://smu.sharepoint.com/sites/oasis/SitePages/DOS-WKLSWC/UCSC.aspx.

Copyright Notice

Please note that all course materials are meant for personal use only, namely, for the purposes of teaching, studying and research. You are strictly not permitted to make copies of or print additional copies or distribute such copies of the course materials or any parts thereof, for commercial gain or exchange.

For the full copyright notice, please visit https://researchguides.smu.edu.sg/copyright.

Accessibility

SMU strives to make learning experiences accessible for all. If you anticipate or experience physical or academic barriers due to disability, please let me know immediately. You are also welcome to contact the university's accessibility services team if you have questions or concerns about academic provisions: accessibility@smu.edu.sg. Please be aware that the accessible tables in our seminar room should remain available for students who require them.

SMU Classification: Restricted

LESSON PLAN		
Lesson	**Reading**	**Topics**
1	Ch 2	Introduction to Python for data analysis
2	TBC	Review of Statistics and Probability ○ Sample vs. population ○ Statistics vs. parameters ○ Descriptive statistics vs. inferential statistics ○ Discrete vs. continuous random variables
3	TBC	Review of inferential statistics ○ Confidence interval estimation ○ Hypothesis testing ○ Bootstrap inference
4	Ch 3	Review Simple Linear Regression ○ Least squares estimation ○ Goodness-of-fit and related inference ○ Gauss-Markov conditions ○ Residual analysis ○ Confidence intervals and prediction intervals
5	Ch 3	Multiple Linear Regression ○ Least squares estimation ○ Model interpretation ○ Qualitative predictors Extensions of the linear model ○ Polynomial regression ○ Transformations ○ Interactive Effects
6		Guest lecture by invited speaker/Industrial visit (TBC)
7	Ch 4	Group Project Presentations 1 Logistic Regression ○ Estimation, prediction, interpretation ○ Confusion matrix, AUC ○ Multiple logistic regression
8	TBC	Time Series Models ○ Serial correlation: ACF, DW Test ○ AR/MA/ARMA models

SMU Classification: Restricted

9	Ch 5	Resampling Methods ○ Bias-variance trade-off ○ Training MSE vs. test MSE ○ Cross-validation
10	Ch 6	Linear Model Selection • Best subset selection • Stepwise selection
11	Ch 6	Linear Model Selection ○ Ridge regression ○ LASSO ○ Elastic Net regression
12		Group Project Presentation 2 Exam