Syllabus for MA 5013
Applied Regression Analysis, Fall 2025

Instructor Information

Instructor:	Rakhi Singh
Class Meeting:	**Thurs @2**, and **Fri @3:30**
Class Location:	KCB 506
Office:	KCB 642
E-mail:	rakhi@smail.iitm.ac.in

Course Information

Description.

This course will cover basic knowledge of linear regression models, including estimation, statistical inference, prediction, model diagnosis, model selection, etc. While the emphasis will be on understanding and assessing of theoretical concepts in the course, we will also delve into computational aspects to see the implementations and the output interpretation in R. Time permitting, we aim to cover the following topics (not necessarily in the same order):

- Simple and multiple linear regression, including polynomial regression.
- Test of significance and confidence intervals for parameters.
- Residuals and their analysis for test of departure from the assumptions such as fitness of model, normality, homogeneity of variances, detection of outliers.
- Influential observations, power transformation of dependent and independent variables.
- Problem of multicollinearity, ridge and principal component regression.
- Model selection of explanatory variables, Mallow's Cp statistic.
- Nonlinear regression, different methods for estimation (Least squares and Maximum likelihood), Asymptotic properties of estimators.
- Generalised Linear Models, Analysis of binary and grouped data using logistic and log-linear models.

By the end of the course, you should expect to

- appreciate and understand the role of statistics in the field of study of your interest.

- know the basic theory of linear regression models: estimation, statistical inference, prediction, model diagnosis, model selection, etc.
- fit and interpret regression models and apply them in various fields of engineering.
- proficient in using programming language R with applications to regression models.
- have basic training in scientific writing and presentation skills.

This course will be adapted from the online course at Penn State's webpage available here: `https://online.stat.psu.edu/statprogram/stat500`

Textbook(s)

I will primarily be referring to the following two books:

Introduction to Linear Regression Analysis by Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining (Wiley), Low price Indian edition is available.

Applied Linear Statistical Models, 5th ed., Kutner et al.,Irwin.

Linear Models with R, Second Edition, Faraway (2014). (Chapman & Hall/CRC Texts in Statistical Science); you can find the R scripts of this book from `https://people.bath.ac.uk/jjf23/LMR/scripts2/`

You can also use other books. While you are not required to buy these books for the class, it is advisable that you read at least one of the books cover to cover. I will also be occasionally referring to:

Applied Regression Analysis by Norman R. Draper, Harry Smith (Wiley), Low price Indian edition is available.

Linear Model Methodology by Andre I. Khuri (CRC Press, 2010)

Some knowledge of the statistical software package R is required. Some parts of the homeworks will typically be in R. There are many online resources where you can learn the basics of R. For example,

- An Introduction to R (`https://cran.r-project.org/doc/manuals/R-intro.pdf`);
- R tutorial by Kelly Black (`https://www.cyclismo.org/tutorial/R/`);
- The undergraduate guide to R by Trevor Martin (`https://biostat.jhsph.edu/~ajaffe/docs/undergradguidetoR.pdf`); and
- a pointer to R bloggers (`https://www.r-bloggers.com/`).

Course Plan

Time permitting, this is the plan for the semester:

Lesson 0	Overview	Week 1
Lesson 1	Simple Linear Regression	Week 1
Lesson 2	Inferences and Prediction in Simple Linear Regression	Week 2
Lesson 3	Multiple Regression	Week 3-4
Lesson 4	Categorical Predictors	Week 5
Lesson 5	Multicollinearity	Week 5
Lesson 6	Prediction and some diagnostics	Week 6
Lesson 7	Regression Diagnostics	Week 6-7
Lesson 8	Transformations, Piecewise Estimation, and GAM	Week 7-8
Lesson 9	Model Selection	Week 9
Lesson 10	Weighted, Generalized Least Squares, and Robust Regression	Week 10
Lesson 11	Penalized Regression	Week 10
Lesson 12	Logistic and Poisson Regression Models and GLMs	Week 11-12
Lesson 13	Tree-based methods	time permitting

Grading

Your grade will be based on your performance on

- quiz 1 (15 percent),
- quiz 2 (20 percent),
- a final project including a report and presentation (15 percent),
- final exam (50 percent),

Ungraded homeworks and their solutions will be provided.

Final Project: Towards the end of the semester, each student will work on a group project (in groups of at most 2). Each student will be using a different dataset to employ regression tools used in the analysis. Each student should have a clear and separate role when participating in the project. This role should be communicated to the instructor. The total points of the project is 100, which will be divided into three parts:

- Initial report (5 pts): due October 3, 2025. The initial report should give description of the data, potential research questions and possible methods to use. The initial report should not exceed one page. The initial report should be sent by the leader of each team to me.

- Presentation (55 pts): each team will give a 5-minute presentation about the outcome of the project. These will likely happen during the final few lecture hours. Depending on the syllabus, we may do this outside class hours.
- Final report (40 pts): due by Nov 7, 2025. The final report should have 4 to 6 pages and should include:
  - Description of research questions / issues (either scientific or statistical question). The significance of the problem you are studying.
  - Description of the data.
  - Preliminary studies: data visualization, dimension reduction, feature extraction, feature selection, statistical inference, model assumption checking (normality? transformation needed?), etc.
  - Statistical analysis
    - Methods: what analyses were done and why. If there is any challenge in analysis, describe your approach to tackle the problem.
    - Results: A small number of well-designed and tailored tables and graphics may be appropriate. No copy-paste of large chunks of software outputs!
    - Conclusion: Convey your findings to broader audience. Discuss any boarder impact.
  - Enclose all your computer code in an individual message to me. No code should be included in the final report.
  - Typos and grammatical errors will be harshly penalized. If you are not yet a master of writing, read The Elements of Style. There are a few copies in the library.
  - The final report should be written with the assumption that the audience of the report are college-educated persons who have taken only elementary statistics. You are NOT writing a report for your professor to read.
  - The final report should not exceed 6 pages, including figures and tables, and must begin with an appropriate title highlighting your choice of topic and analysis.
  - Data Sources. Find your own data set online (e.g. google "predictive analytics data set"), you will find plenty. Some data repositories are: `https://www.kaggle.com/datasets`,Ψ`https://archive.ics.uci.edu/ml/index.php`.

Exams: The final end-semester exam will be comprehensive. Quizzes 1 and 2 will be held in the institute-planned slots. The material covered in quizzes will remain flexible and will be conveyed in due course of time.