Imperial College
London

Advanced Regression Course Handbook 2016-2017

Module Leader:
Dr Korbinian Strimmer

Advanced Regression

TIMETABLE

Date	Time	Session Type	Session Title	Lecturer
16 Jan - Course Week 16	10.00-10.45	Lecture 1a	Overview and motivation	Korbinian Strimmer
	11.00-12.00	Lecture 1b	Statistical learning using likelihood	Korbinian Strimmer
	1.30-2.15	Lecture 1c	Simple linear models	Korbinian Strimmer
	2:30-4:30	Practical 1	Using R to analyse data with linear models	Tutors
23 Jan – Course Week 17	10.00-11.00	Lecture 2a	Introduction to variable selection	Korbinian Strimmer
	11:15-12:00	Lecture 2b	Variable selection with correlated predictors	Korbinian Strimmer
	1:30-2:15	Lecture 2c	Prediction accuracy and cross-validation	Korbinian Strimmer
	2:30-4:30	Practical 2	Cross-validation and variable selection in R	Tutors
30 Jan – Course Week 18	10.00-10.45	Lecture 3a	Challenges in high-dimensional data analysis	Korbinian Strimmer
	11.00-12.00	Lecture 3b	Penalised regression models (lasso, elastic net)	Korbinian Strimmer
	1:30-3:30	Practical 3	High-dimensional regression in R	Tutors

6 Feb – Course Week 19	10.00-11.00	Lecture 4a	Nonlinear and nonparametric models - part I (lowess, spline, GAM)	Korbinian Strimmer
	11.15-12.00	Lecture 4b	Nonlinear and nonparametric models - part II (decision trees and random forests)	Korbinian Strimmer
	1:30-3:30	Practical 4	Nonlinear and nonparametric models in R	Tutors

13 Feb – Course Week 20	10.00-11.00	Lecture 5	Random effects and hierarchical models	Korbinian Strimmer
	11:15-12:00		Questions and answers session	Korbinian Strimmer
	1:30-3:30	Practical 10	Random effects models in R	Tutors

COURSE OUTLINE

MODULE LEADER:

Dr Korbinian Strimmer
E-mail: k.strimmer@imperial.ac.uk

MODULE STRUCTURE

The "Advanced Regression" module is a an optional module for the MSc in Epidemiology. It will run over the first 5 weeks of Term 2. Sessions will take place on Tuesdays for the whole day, and will be a mix of lectures and computer-based practicals.

The practicals are an essential part of the course, where students will learn how to use the statistical analysis software R for advanced regression analysis. The practicals also serve as opportunities to explore the theoretical concepts introduced in the lectures, ask questions and enable students to acquire the skills necessary to perform advanced statistical data analysis during the mini projects or MSc dissertation.

The module "Advanced Regression" will cover topics such as high-dimensional data analysis, ridge and lasso regression, variable selection and statistical analysis using random forests. The relationship to Bayesian statistics is also discussed, as most techniques for advanced regression discussed in this module may be viewed as computationally efficient approximation to full Bayesian modeling.

ASSESSMENT AND FORMATIVE FEEDBACK

This module will be assessed formally in the Paper 2 examination in May 2017.

Practical sessions always offer great opportunity for receiving formative feedback from the tutors. Two revision sessions, one on the last day of the module and another just prior to the exam, will be scheduled.

COURSE AIMS

The aim of the module is to ensure that the students familiarise with the principles of advanced regression for high-dimensional data so that they are able to apply such techniques on real data problems (e.g. complex omics data). They will also further familiarise with R, the software to implement the models presented in the module.

Knowledge in advanced statistical regression methods will be highly useful for carrying out the mini-projects and the summer dissertation.

Core concepts:

- Principles of likelihood inference
- Overview of generalized linear models
- Evaluation of model performance (e.g. predictive accuracy)
- Resampling methods such as cross-validation and bootstrap
- Methods for variable ranking and selection as well as for model selection
- Challenges in high-dimensional data analysis
- Penalised regression models: ridge regression, lasso, elastic net etc.
- Shrinkage estimation
- Nonlinear and nonparametric regression models
- Random effects models and hierarchical models
- Conceptual connections with Bayesian paradigm
- R: how to use the software to perform advanced regression analysis

LEARNING OUTCOMES

By the end of this module, students should be able to:

- Perform advanced statistical analyses, employing penalised likelihood or nonpararametric regression models
- Understand the theoretical foundations and limitations of the most widely used advanced regression approaches
- Recognise the challenges of high-dimensional data analysis
- Know suitable analysis strategies to address the problems arising from “small n, large d” data sets.
- Understand the close relationship among Bayesian, likelihood, and penalised likelihood inference and shrinkage methods.
- Practically build and employ complex regression models in R

RECOMMENDED GENERAL BACKGROUND READING

James, Witten, Hastie and Tibshirani. 2013. *An Introduction to Statistical Learning with Applications in R*. Springer.
PDF available from http://www-bcf.usc.edu/~gareth/ISL/

Wood. 2014. *Core Statistics*. Cambridge University Press.
PDF available from http://www.maths.bris.ac.uk/~sw15190/core-statistics.pdf

ON-LINE RESOURCES

R Analysis Platform
http://r-project.org

SYNOPIS OF LECTURES

Course Week 16:

LECTURE 1a Overview and motivation

Learning Objectives

After this session students should be able to:

- understand why high-dimensional and other advanced regression approaches are important in modern biomedical statistical data analysis
- understand key differences of advanced regression methods in comparison with the classical statistical methods introduced in Term 1

LECTURE 1b Statistical learning using maximum likelihood

Learning Objectives

After this session students should be able to:

- understand the basic principles of maximum likelihood inference
- analytically derive some simple point maximum likelihood point estimates
- compute error bounds using Fisher information
- know about optimality properties of maximum likelihood in the large sample domain

LECTURE 1c Simple linear models

Learning Objectives

After this session students should be able to:

- understand the workings of both linear regression models and generalisations such as logistic regression
- know essential maximum likelihood estimates for linear models
- understand generalized linear model (GLM) terminology
- employ and interprete diagnostic checks of model fit
- practically apply these models to actual data using R

Course Week 17:

LECTURE 2a Introduction to variable selection

Learning Objectives

After this session students should be able to:

- understand the importance and challenges of variable selection in statistical modeling

- know classical procedures for variable selection (incl. AIC and AICc based on prediction and BIC) and comparing models (e.g. Kullback-Leibler information)
- understand the connections between variable ranking, variable importance and variable selection procedures
- practically be able to use variable selection in R

LECTURE 2b Variable selection with correlated predictors

Learning Objectives

After this session students should be able to:

- understand the problems that highly collinear predictors pose for regression analysis
- know strategies to circumvent collinearity, e.g. by pre-processing the data to group variables or to whiten the data.
- Adddress problems of collinearity in practical data analysis in R.

LECTURE 2c Prediction Accuracy and cross-validation

Learning Objectives

After this session students should be able to:

- understand concept of prediction accuracy and related quantities
- understand various resampling procedures (including cross-validation) to estimate prediction error and suitable to choice of parameters (e.g. folds)
- practically use cross-validation procedures in R

Course Week 18:

LECTURE 3a Challenges in high-dimensional data analysis

Learning Objectives

After this session students should be able to:

- understand challenges of data analysis when the dimension of a model is equal or exceeds the number of data points ("small n, large d" setting)
- understand failure of maximum likelihood estimates in this settings (e.g. overfitting, singularity, instability)
- know strategies to overcome limitation due to dimensionality (regularisation, penalisation, adding prior information, sharing information, shrinkage)
- know simple shrinkage estimators dominating the maximum likelihood estimator (James-Stein estimator, for mean, variance and correlation)

LECTURE 3b Penalised regression models (ridge, lasso, elastic net)

Learning Objectives

After this session students should be able to:

- understand the basic three regularized regression models (ridge regression, lasso regression, elastic net)
- be able to use these methods in R

Course Week 19:

LECTURE 4a Nonlinear and nonparametric models – part I (lowess, spline, GAM)

Learning Objectives

After this session students should be able to:

- understand smoothing based regression models (lowess, spline)
- understand the concepts of generalized additive models (GAM)
- apply these models in R

LECTURE 4b Nonlinear and nonparametric models – part II (decision trees and random forests)

Learning Objectives

After this session students should be able to:

- understand the basic concepts of decision trees and random forests
- practically apply these models to actual data using R

Course Week 20:

LECTURE 5 Random effects and hierachical models

Learning Objectives

After this session students should be able to:

- understand the principle behind random effects regression models, and by extension of mixed effects regression models
- be able to analyse data in R with random/mixed effects linear models