Examinable Material: Linear and Generalised Linear Regression Models

General information

- Answer 3 questions from a choice of 4.
- The 4 questions set will be a random selection of 4 from the 5 question types discussed below. Please note carefully: the demarcations between topics in the question types in this document are for guideline purposes and are not strict. In practice this means that a modest amount of overlap between coverage among the question types outlined below is possible.
- All questions carry equal marks and have equal difficulty levels.
- 2 hours – so 40 mins a question.
- Make sure that you can get 40+% comfortably from each question chosen.
- Move on after allowed minutes – do NOT get bogged down – just leave it.
- Any queries - get the Examinations Office to contact me and do something else in the meantime.
- You can have a copy of the NCST tables if you like.
- You are allowed use your own non-programmable calculator – check your model is on the list maintained by the Examinations Office.

Q. Type 1: Multiple Regression Model and Least Squares - with a focus on models consisting of continuous predictors

- The least squares criterion - what is is? Can you write a mathematical expression for it in general or for a particular case?
- Interpretation of the slope parameters and intercept. Role of the intercept. T-tests for individual parameters (variance-covariance matrix of parameters).
- Type I and II errors and the multiple testing problem. The purpose and logic of the F-statistic. Decomposition of the total sum of squares (formula) into the model SS and error SS (formulae).
- The ANOVA table, null and alternative hypotheses; R^2 and its interpretation; recovering the complete ANOVA table from the default `lm(.)` output.
- R code for fitting such models including the `lm(.), update(.) and drop1(.)` functions.
- Predicted (aka fitted) values; Confidence intervals and prediction intervals (compare and contrast these?). R code for calculating these.

- Geometric interpretation for 2–D and 3–D models (degree 1 and 2 polynomials in 2 and 3 dimensions).
- General linear hypotheses for testing customised hypotheses, the L matrix/vector etc. R code for doing these e.g. using the `multcomp` library and the `glht(.)` function.

See: Lectures(s) 1–3 and associated materials; R classes/code; essential reading; past papers.

Q. Type 2: Multiple regression models including categorical predictors.

- Including categorical predictors using dummy variables (set-to-zero constraint) – how and why are these used?
- Default dummy variable coding in R using the `factor(.)` function.
- Interpreting parameters for dummy variables.
- Geometric interpretation of models with a categorical and a continuous predictor with/without interaction.
- Interactions of two categorical predictors - purpose and interpretation of such interactions – code for fitting such interactions – hypothesis testing of such interactions.
- General linear hypotheses for testing customised hypotheses involving categorical predictors and interactions, the L matrix/vector etc. R code for doing these e.g. using the `multcomp` library and the `glht(.)` function.
- Pairwise comparisons of treatment means. Calculating treatment least-squares means and differences between means from the regression parameters. Experimentwise error rates for pairwise comparisons (including the number of comparisons formula). Methods for correcting p-values applied to pairwise comparisons: None; FDR; Holm and Bonferroni methods.

See: Lectures(s) 3–4 and associated materials; R classes/code essential reading ; past papers.

Q. Type 3: Model building issues:

- Formula for calculating how many possible different models may be fitted with k predictors.
- Descriptions of and problems caused by under-fitting and over-fitting.
- Be able to define the concept of best model (1 line).
- R^2, adjusted R^2, all possible regressions. Weakness of using these on their own for model building.
- Akaike information Criterion (AIC) and its uses in model identification.
- Describe in detail how forward, backward and stepwise selection procedures work using both p-values and the AIC. Logic for using higher than optimal α level.
- If presented with a series of outputs using the `add1(.)` and `drop1(.)` functions be able to implement the forward, backward or stepwise procedures using hypothesis testing.

- If presented with a table with log likelihoods be able to implement the forward, backward or stepwise procedures using the AIC. R Code for applying the procedures; `step(.)` function and `scope(.)` argument.
- The LASSO algorithm – the LASSO penalised LS criterion – the role played by the shrinkage parameter λ. Estimating λ using cross validation - outline of the steps involved. Implementation of the the procedure using the `glmnet` library.

See: Lectures(s) 5 and associated materials; R classes/code; essential reading ; past papers.

Q. Type 4: Model Diagnostics:

- Define raw residuals and explain how they are calculated. Plots of raw residual and the use of such plots in detecting model misspecification.
- Expected value and variance of the residuals. Externally and internally studentised residuals their uses in outlier detection. Hypothesis testing on residuals - Bonferroni correction, possible null and alternative hypotheses.
- Description and uses of QQ plots.
- Discuss influence measures: dffits; dfbetas and Cook's D.
- Multicollinearity - what is it and the problems it can cause.
- Spotting multicollinearity; correlation matrix and correlation lattice plot; Variance inflation factors - how these are calculated and interpreted.

See: Lectures(s) 6 and associated materials; R classes/code; essential reading; past papers.

Q. Type 5: GLM - Logistics & Poisson regression

- Model formulation for logistic regression fitted to a binary response or Poisson regression fitted to a count response.
- The role of the linear predictor (η_i) and relating the linear predictor to the response using the inverse logit function or log function.
- Likelihood for both models.
- The maximum likelihood method. Fitting the models using R and the `glm(.)` function.
- Interpret model parameters and calculate log odds, odds, log odds ratios, odds ratios, probabilities. How to calculate CI's for all of these and the R code for doing so. Calculate expected counts and/or rates from Poisson regression models - CIs for these.
- Hypothesis testing: Wald and LR based methods - describe the both and apply both given appropriate output from R . Custmised hypotheses using the `multcomp` library.
- Offsets and their use in Poisson regression.

See: Lectures(s) 7 & 8 and associated materials; R classes/code; essential reading; past papers.