MTHM506 Statistical Data Modelling

Problem Sheet 2
(Covers Topics 3-4)

You should attempt all questions on this sheet. The questions constitute both summative (indicated by marks) and formative assessment. Marks achieved in this assignment will contribute 25% of the final module mark. ***Solutions are expected to be clearly explained, concise, well structured and well presented. Give R input commands for each model fitted (e.g. 'model <- gam(...)'). Do not display too much raw R output as part of your solutions (e.g. don't display the full output of 'summary(model)'), but edit this down to the essentials. All plots should have titles and appropriately labelled axes. Hand written solutions will be accepted, but a more professional word processed submission is preferred (as is a mixture of the two).***

Topic 3

1. The data set `globalMeanTemp` contains global mean temperature monthly "anomalies" for 1880–2013. The term "anomaly" means an overall mean value was substracted from the original data points. Interest lies in quantifying the temporal trend (change in the mean over time) in the data.

   (a) Why does an additive model make sense for this data set?

   (b) Fit an additive model to this data, ensuring this has adequate flexibility and that it fits the data well.

   (c) Produce a plot of the data and estimated mean (trend) along with 95% confidence as well as prediction intervals.

   (d) Investigate (using additive models) whether there is a significant seasonal cycle in the data.

2. The dataframe `carbonD` contains monthly observations of CO2 concentration (in parts per million) from 1959 to 1997, measured at Mauna Loa (Hawaii).

   (a) Plot `co2` against `timeStep` and use this to suggest and write down a plausible GAM to describe this data set. (6)

   (b) Fit the suggested GAM ensuring the fit is good. (4)

   (c) Use the function `predict` with argument `type="terms"` to plot estimates of any smooth functions in your model. (2)

   (d) Use the model to predict year CO2 for the year 1998 and produce a plot of this along with 95% prediction intervals. (3)

   (15)

3. In this question we return to the AIDS data from sheet 1: the number of quarterly aids cases in the UK, y_i, from January 1983 to March 1994.

   (a) Fit a GAM to this data to describe the number of AIDS cases as a function of time. (7)

   (b) Plot the estimaed mean from the model as well as 95% confidence and prediction intervals. (4)

   (11)

4. The dataframe `munichrent` which contains data on net rent per month in Euro ('rent') for more 3,000 randomly sampled similar sized apartments in Munich Germany. Other variables in the data set are the year of construction ('yearc') the quality of the location according to an expert assessment (a three level factor ('location') — average location (1), good location (2), top location (3)) and an additional factor ('district') giving the actual district of munich in which the apartment is situated. Interest lies in the relationship (if any) between net rent and the year of construction and also the expert assessment of the location and possibly the interaction between year of construction and assessed quality of location.

   (a) Fit a Normal GLM to the relationship between net rent and year of construction, location assessment and the interaction between them. Produce a density plot of the distribution of the residuals and a Q-Q plot of the residuals and also plot the residuals against the fitted values from the model. On the basis of these plots explain with reasons why the linear model assumptions of a Normal distribution with constant variance are not suitable for this data set.

(b) Suggest an alternative GLM for this data set using the same predictors and choosing a suitable distribution from the exponential family with an appropriate link function. Justify your choices. Fit your proposed model, report and interpret the results and carry out appropriate model checking.

(c) Extend your GLM to an equivalent GAM where the linear relationship between net rent and year of construction and it's interaction with assessed quality of location are replaced by appropriate smooth functions. Fit this model, report and interpret the results, plot the associated estimated smooth functions, and carry out appropriate model checking. Say (with reasons) whether you prefer the GAM formulation to that of the GLM.

Topic 4

5. A tyre manufacturer wants to investigate amount of tyre wear, and whether this differs substantially among the four positions on a car that each tyre can occupy (front offside, front nearside, rear offside and rear nearside), and also whether the car type affects the amount of wear (i.e. whether there is an interaction). An experiment was conducted in which tyres were fitted to a car, the car was driven at a fixed speed and distance and the reduction in depth of tread caused by the test was measured (in hundredths of a mm) for each tyre. The process was repeated 3 times with a new set of tyres each time, for 3 different types of car (A, B and C). Results obtained are given in the data frame `tyres`.

   (a) Fit a Normal GLM to test the effects of tyre position and car type.

   (b) Briefly explain why it would be sensible to consider the effects for `car_type` as random instead of assuming they are fixed, and fit a model with wheel position as a fixed effect and car type as a random effect. Write down the mathematical formulation of this model and the estimates of the conditional variance of the response and the variance of the random effects.

   (c) Test for the significance of the fixed effects, using both the likelihood ratio test and a bootstrap test, stating which is preferable and why.

   (d) Repeat this to test for the significance of the random effects.

   (e) Use the standard errors produced by `R` to conduce crude significance tests for the parameters of car type. State any assumptions you are making.

   (f) Conduct more accurate significance tests using bootstapping to construct confidence intervals.

   (g) Predict the mean wear for the front offside wheel for car A but also the mean wear for an entirely new car.

   (h) Produce the residuals vs fitted values but also the QQ plot of the residuals and comment of the validity of the model.

6. An experiment was conducted to compare 4 different treatments (A,B,C, and D) on the production of penicillin. The material used for producing the penicillin is quite variable and it can only be made in blends sufficient for 4 runs. So the data consists of 5 blends and the 4 treatments were applied to each blend. The data can be found in dataframe `penicillin`. Clearly, the treatment can be considered as fixed effects, however blend should be random (many more blends could have been chosen).

   (a) Fit a normal mixed model with a fixed effect for treatment and a random effect for blend, making sure you define the model mathematically. By fitting appropriate reduced models, test for the significance of both the fixed and the random effects using likelihood ratio tests. (10)

   (b) Comment on the validity of using these tests in mixed effects models, suggest an alternative way of implementing these tests, and use it to compare with results in (a). (5)

   (15)

7. The dataframe `pupils` which involves language scores in Dutch schools. This is an example of a two level situation. Specifically, the data considers 131 schools (but only 1 class per school) for $i = 1, 2, \ldots, 2287$ students in grades 7 and 8. The nesting therefore occurs within each school $j = 1, \ldots, 131$.

   Interest lies in assessing the impact on language scores of pupil factors such as IQ (`IQ`) and pupil social status (`ses`). The response variable is denoted as `test`. The categorical variable (factor) `Class` refers to the class that each pupil belongs to (so `Class`$= 1, \ldots, 131$).

(a) First fit a (Normal) linear model using `glm()` with `test` as the response and `IQ`, `ses` and factor `Class` as the covariates. Comment on the significance of the two continuous variables and perform a likelihood ratio test to test on the overall significance of the factor `Class`. (2)

(b) i. State two reasons why one might want to treat the class effects as random. (4)

ii. Write down the mathematical formulation of a Normal random effect model (`IQ` and `ses` as fixed effects and `Class` as a random effect). (2)

iii. Fit this model and comment on the significance of the fixed effects based on the t-tests. State any assumptions you are making. (3)

iv. What is the estimate of the "within-class" variance and the "between-class" variance. What is the estimate of marginal variance of the response, it is different to the (marginal) variance from the model in (a) and if so, why? (5)

v. Test whether the variance of the random effects is zero, using a likelihood ratio test and comment on its validity. (6)

vi. Plot a density estimate of the predicted random effects and superimpose their theoretical Normal distribution using the estimate of their variance. Use functions `qqnorm()` and `qqline` to produce a QQ plot of the random effects and comment on the validity of a Gaussian model for the random effects. (4)

vii. Note that the functions `fitted()` and `resid()` in `lme4`, will produce the fitted values $\hat{y}$ and raw residuals $y - \hat{y}$. Use these functions, in conjunction with the two functions in (iii) to produce a QQ plot of the residuals and a residuals vs fitted values plot. Comment on the model assumptions using the two plots. (4)

(c) One of the student-level covariates is IQ which may affect the test results per student. However, there may be class level (latent) variables, such as teacher competence, which may have an effect on how IQ relates to the test result in each class. Such a scenario may be accommodated by considering the parameter of `IQ` to be random rather than it being fixed (and constant across classes).

i. Extend the model in (b) to make the parameter of `IQ` vary with `Class`. Compare (qualitatively) the overall effect of `IQ` on `test` between this model and the model in (b). (4)

ii. Test for the significance of the random slope for `IQ` using a likelihood ratio test. (3)

(37)

8. In this question we return to the dataframe `sexr` which relates to the ratio of male to female births across various countries. This can of course be modelled as a Binomial situation with number of male births (`Male`) out a total number of births (`Total`). The hypothesis is that the male to female birth ratio decreases when "things are bad". As such, two deprivation measures are included: `totleatbirth` (life expectancy at birth) and `tinmort` (infant mortality) – higher value of which indicates more deprivation. There is also information on each country (`Country`) to allow for country variability of the ratio.

(a) Fit a Binomial GLM with `totleatbirth`, `tinmort` and `Country` as covariates. Check whether the model fits (w.r.t. to the saturated model) and also produce the QQ plot of the residuals.

(b) Now fit the model as a Binomial GLMM with a random effect for each observation. State why we would not have to check whether this model fits (w.r.t. to the saturated model), and also produce the QQ plot of the residuals and compare with the GLM.

(c) Check for the significance of the random effects using the likelihood ratio test, commenting on its reliability.

(d) Comment on the significance of the fixed effects and they say about male to female birth ratio.

9. The dataframe `hip` contains information on the number of hip fractures (`Nfract`) in elderly population per municipality (`municipality`), in Portugal. The data are stratified by gender (`sex`, 1=males and 2=females) and the socio-economic status, `ses`, of each municipality (factor 1=poor, 2=average, 3=good). Interest here lies on the effect of `ses` on the rate of hip fractures per 1000 population, after allowing for gender. The data frame also contains information on the number of people (`Npop`) in 1000s, in each `sex` and `ses` combination per municipality.

(a) Give two reasons why we would want to treat the municipality effects as random. (4)

(b) Write down (mathematically) an appropriate GLMM with an appropriate offset to assess the effect of `ses` on the hip incidence rate. (6)

(c) Fit this model in R and comment on parameter significance using the z-tests. (3)

(d) Produce a QQ plot of the residuals and comment appropriately. Also produce a QQ plot of the random effects and comment on the appropriateness of the Normal distribution. (4)

(e) The function `glmer` fits the model using maximum likelihood. Use this fact to perform a likelihood ratio test to assess the significance of the random effects, commenting appropriately on the validity of the test. (5)

(22)

(Total=100)