MATH70071 Applied Statistics – Assessed Coursework 3

Due Friday 10 December 2021, 6pm

Submit no more than **8 pages**. Keep any R code concise and present it inline as part of the report. Considerable emphasis will be put on clarity of expression and a clean presentation. Only detailed, well-written answers will score highly.

Question 1

Recall from Assessed Coursework 2 we fitted a generalized linear model to count data from the R file glmxy.R posted on Blackboard, which contained a data frame called `dfrm` with 500 observed counts (`dfrm$y`) and a single associated predictor (`dfrm$x`) associated with each observed count. The data were later revealed to have been generated from a negative-binomial regression model with $r = 2$, using a (canonical) logarithmic link function of the form $\eta = -\log(1 + r/\mu)$, and intercept and slope parameters $\beta = (-0.5, -2.5)$.

a) Fit a Poisson regression GLM to these data using the canonical link function (denoted g), assuming the presence of an intercept term. Report the fitted regression coefficients $\hat{\beta}$.
b) Using the fit from part a), plot the response values y_i against the inverse link function of the estimated linear predictors, $g^{-1}(\hat{\eta}_i)$. Draw the fitted regression line on this plot, and comment on the apparent goodness of fit.
c) Add another (dashed) line showing the true regression function on these axes. (You can just include one plot in your submission.) Compare and comment on the difference between the two regression lines.
d) Under the fitted Poisson regression model from part a), calculate an approximate 99% confidence interval for the mean response value for the covariate value $x = 0.5$.
e) Plot the deviance residuals d_i against the predictor values x_i. What proportion of these residuals are negative?
f) Suppose now that we are interested in modelling the conditional distribution of x given y. Fit a linear model for x with the variable y treated as a factor variable. Explain the magnitude of the p-values observed in the summary, and comment on the regression coefficient estimates corresponding to each level of y. Fit the same model a second time treating the factors as random effects. Compare the estimated coefficients obtained under these two approaches. Considering the differing the assumptions of the two approaches, which estimated coefficients are you more inclined to trust?

25 Marks

Question 2

The R file xyz.R posted on Blackboard contains a data frame called `xyz` with 480 observations of a continuous valued response variable y, with an associated integer-valued predictor (x) and a category label A–H (z) indicating membership of the observation to one of eight groups. Load the data using the command `source("xyz.R")`.

a) Make scatter plot of the response variable y against the predictor variable x, using a different mark/symbol for each category A–H. Fit a normal linear model for y against the single predictor x with the inclusion of a global intercept term, and add the fitted regression line to the plot. From this plot, comment on two aspects of the data which make this simple linear model appear to be inappropriate.
b) Make box plots of the response variable y for each z category, commenting on the between category comparison. Calculate the sample mean response value for each group.
c) Assume a normal linear mixed model for y assuming a fixed effect for x together with an intercept, and random effect intercept terms corresponding to membership of the z categories A–H. Report restricted maximum likelihood estimated values for the error variance and the variance of the random effects.
d) Now remove observations from categories A, C, D, E and G from the data. Calculate the value of the maximised unrestricted log-likelihood function from the normal linear mixed model on this reduced data set.
e) On the same reduced data set, fit a simple linear model for y by excluding the random effects. Report the maximised log-likelihood, and the deviance between the two models (with and without inclusion of the random effects).
f) Use the parametric bootstrap procedure to investigate the significance of the category labels in the mixed linear model with the reduced data set, reporting an estimated p-value after 1000 simulations. Interpret the p-value, and explain this result.

20 Marks

End