MATH70071 Applied Statistics – Assessed Coursework 2

Due Monday 15 November 2021 – deadline specific to your group

Upload your final version only - once the report is uploaded there is no option for re-uploading. Avoid last minute uploads. Hand-in no more than **10 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 the data from Coursework 1 concerning various factors affecting median house prices in different suburbs of Boston. The data can be obtained from the R file bos.R posted on Blackboard.

In Coursework 1 we fitted a normal linear regression model for nitric oxide concentration (NOX) with the following predictors: INDUS, RAD, TAX, AGE, along with an intercept term. In this coursework we will extend this model by categorising the AGE variable.

a) Make a scatter plot of NOX against AGE. Without performing any analysis, comment on whether a linear relationship for NOX against AGE looks appropriate, either globally or in some AGE interval.
b) Create a factor variable AGEBAND with three levels: "low", "medium" and "high", corresponding to AGE values in (0,25], (25,80] ad (80,100] respectively. How many observations are there in each category?

To augment the previous linear model with this AGEBAND variable, we can introduce the following notation. Respectively denoting the categories "low", "medium" and "high" as 1, 2 and 3, let y_{ij} denote the NOX level for the ith observation in the data from AGEBAND category j, $i = 1,\ldots,n_j$, $j = 1,2,3$, with n_j denoting the number of observations in category j [calculated in part b)].

The revised linear model we now consider is

$$y_{ij} \;=\; \alpha_j \;+\; \sum_{k=1}^{4} \beta_k \, x_{ijk}, \quad i = 1,\ldots,n_j;\, j = 1,2,3, \tag{1}$$

with $\beta \in \mathbb{R}^4$. For observation i from AGEBAND category j, x_{ijk}, $k = 1,2,3,4$, denotes the respective value of the variables INDUS, RAD, TAX, AGE. The remaining terms $\alpha \in \mathbb{R}^3$ represent the main effects for each AGEBAND category. Note that the AGE variable appears twice in this model, once directly and once through a categorical transformation.

c) Suppose we are interested in whether there is a change in NOX levels between the medium and high age categories, when also conditioning on the other variables in the model (1). State the implied parameter contrast of interest, and report the estimate of that contrast, the t-statistic and the significance level p-value.
d) Repeat the analysis of part c) for contrasting the low and medium age categories. Compare your findings with those from part c). Do the results from these two parts agree with what we might have anticipated?
e) Report a 99% confidence interval for α_3, the main effect parameter for the "high" age category.

20 Marks

Question 2

Consider the probability mass function for a non-negative integer y,

$$\mathbb{p}(y \,|\, p) = \binom{y+r-1}{y} \, p^y \, (1-p)^r, \quad y = 0,1,2,\ldots, \tag{2}$$

with $0 < p < 1$ an unknown parameter and $r \geq 0$ assumed known.

a) Show that the probability mass function (2) is from an exponential family, meaning it can be expressed as

$$p(y \,|\, \theta,\phi) = \exp\left\{ \frac{\theta y - b(\theta)}{a(\phi)} + c(y,\phi) \right\}, \tag{3}$$

specifying the natural parameter θ and a sensible choice for the scale parameter ϕ. State equations for the following quantities: $a(\phi), b(\theta), c(y,\phi), b'(\theta), b''(\theta), \mu, b'^{-1}(\mu), V(\mu)$, where μ is the mean and $V(\mu)$ the variance function.
b) Using the results from a), or otherwise, identify the mean $\mathbb{E}(y \,|\, \theta,\phi)$ and variance $\mathbb{V}(y \,|\, \theta,\phi)$. Is the variance smaller or larger than the mean for this distribution?

Continued on next page

c) Express the original parameter p as a function of the mean value μ and the known value r. For a fixed mean μ, find the limiting distribution of (2) as $r \longrightarrow \infty$.

[Hint: For fixed values of r, y, μ, note that $\lim_{r\longrightarrow\infty} \frac{(y+r-1)!}{(r-1)!(\mu+r)^y} = 1$.]

d) Under which practical circumstances might (2) be considered preferable to a Poisson distribution as a statistical model for count data? Include in your answer the investigation of limiting behaviour of (2) as $r \longrightarrow \infty$ from part c).

Consider a vector of n, non-negative integer response variables $\mathbf{y} = (y_1, \ldots, y_n)$ with an $n \times p$ matrix of associated covariates X, with ith row $x_i = (x_{i1}, \ldots, x_{ip}) \in \mathbb{R}^p$. Suppose we wish to fit a generalized linear model for $y \mid X$, using distributions $\mathbb{p}(y \mid p_i)$ of type (2) with observation specific parameters $p_i \in (0,1)$; the parameter p_i for observation i shall be determined by the covariate x_i through the canonical link function applied to the linear predictor

$$\eta_i = \beta \cdot x_i,$$

with $\beta \in \mathbb{R}^p$. [Recall the canonical link function sets $\theta_i = \eta_i$, where θ_i is the natural parameter from (3) for observation i.]

e) Assuming the canonical link function, state the implied equation for linking p_i to η_i. Does this link function present any problems?

f) Still assuming the canonical link function, derive the following quantities (as defined in the lecture notes) required for the iterative weighted least squares (IWLS) algorithm:
   i. μ_i, as a function of η_i
   ii. $\partial \eta_i / \partial \mu_i$
   iii. $\tilde{w}_{ii}$

g) Write R code to perform IWLS estimation of the regression coefficients $\beta \in \mathbb{R}^p$.

h) The R file glmxy.R posted on Blackboard contains a data frame called `dfrm` with 500 observed counts (`dfrm$y`) with a single associated predictor (`dfrm$x`) associated with each observed count. Load the data using the command `source("glmxy.R")`. Use the R code from the previous part to obtain IWLS estimates for the intercept and slope from the assumed generalized linear model when applied to these data.

30 Marks

End