MTHM506 - Statistical Data Modelling

Individual assessment sheet

Marks achieved in this assignment will contribute towards 50% of the final module mark. You should attempt **all** questions on this sheet. Note that the questions are organised in the order we covered the topics, and not in order of difficulty. Therefore it is advised that you read through the questions first, and start working on those that you feel more comfortable with.

Deadline: Noon (12pm), on 28th February 2025

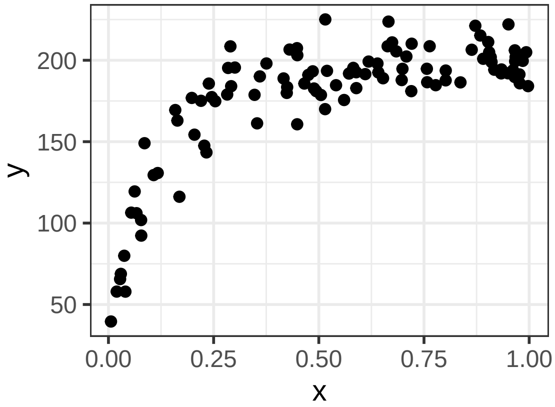
You should submit one pdf **via ELE** containing your solutions - it should be written up using word processing software (e.g. LaTeX, R Markdown, or Word). Solutions are expected to be concise, well structured and well presented. Commented `R` code (e.g. '`model <- glm(...)`') and the outcomes/plots should form part of your solutions. Do not display too much raw `R` output (e.g. don't display the full output of '`summary(model)`'), but edit this down to the essentials. Ensure to include justification for each step of your analyses, providing comments alongside your `R` code to explain what you are doing and add appropriate titles and labelled axes to your plots.

You are expected to work independently - strict disciplinary action will be taken for any plagiarism. Late submissions will also be penalised.

The data required for this assignment are in part of `datasets.RData` which can be downloaded from the ELE page and loaded into `R` using the `load()` function.

Question 1 [25 marks]

The dataframe `nlmodel` contains data on a response variable `y` and a single explanatory variable `x`. A scatter plot of `y` versus `x` suggests a strong non-linear relationship:



Suppose for these data we wish to consider the model

$$Y_i \sim \mathsf{N}\left(\frac{\theta_1 x_i}{\theta_2 + x_i}, \sigma^2\right)$$
$$i = 1, 2, \ldots, 100, \quad Y_i \text{ independent}$$

(a) [*2 marks*] Why can't this model be fit using a linear (regression) model?
(b) [*2 marks*] Write down the likelihood $L(\theta_1, \theta_2, \sigma^2; \boldsymbol{y}, \boldsymbol{x})$ and the log-likelihood $\ell(\theta_1, \theta_2, \sigma^2; \boldsymbol{y}, \boldsymbol{x})$
(c) [*1 mark*] Write an `R` function `mylike()` which evaluates the negative log-likelihood (i.e. $-\ell(\theta_1, \theta_2, \sigma; \boldsymbol{y}, \boldsymbol{x})$) for any values of the three parameters
(d) [*7 marks*] Use the `R` function `nlm()` in association with your function `mylike()` to numerically minimise the log-likelihood. Provide some evidence of how you chose sensible starting values. Report the maximum likelihood estimates of the parameters and superimpose a plot of the associated mean relationship on a scatter plot of `y` versus `x`.
(e) [*6 marks*] Report the standard errors for θ_1 and θ_2, and use those to construct 95% confidence intervals.
(f) [*3 marks*] Test the hypothesis that $\theta_2 = 0.08$ at the 5% significance level (not using the confidence interval) and compute the associated p-value of the test.
(g) [*4 marks*] Use plug-in prediction to construct and plot 95% prediction intervals.

Question 2 [30 marks]

The dataframe `aids` data relates to the number of quarterly AIDS cases in the UK, y_i, from January 1983 to March 1994. The variable `cases` is y_i and `date` is time, symbolised here as x_i. In this question we consider two competing models to describe the trend in the number of cases. Model 1 is

$$Y_i \sim \mathsf{Pois}(\lambda_i)$$
$$\log(\lambda_i) = \beta_0 + \beta_1 x_i$$

and Model 2 is

$$Y_i \sim \mathsf{N}(\mu_i, \sigma^2)$$
$$\log(\mu_i) = \gamma_0 + \gamma_1 x_i$$

(a) [*3 marks*] Plot y_i against x_i and comment on whether the two proposed models are sensible in terms of the distribution and the relationship of x with the mean.
(b) [*5 marks*] Fit the two models in `R`. Plot the estimated trends from each model ($\hat{\lambda}_i$ and $\hat{\mu}_i$) on top of the data with approximate 95% confidence intervals around the mean. Comment on the validity of each model (based on the plot). Obtain the AIC for each model and thus comment on which model is preferable.
(c) [*3 marks*] Produce the deviance residuals vs fitted values ($\hat{\lambda}_i$ and $\hat{\mu}_i$) plot for each model, comment appropriately and thus propose a way that the two models might be extended to improve the fit.

(d) [*4 marks*] Implement the proposed extensions to each model, to arrive at a final version for each of them (justified by appropriate hypothesis tests).

(e) [*11 marks*] On the basis of your answer to (a), analogous plots as in (b) and (c), but also on arguments of model fit based on the deviance and the AIC, comment on which (if any) of the two final models in (d) you would choose as the best. Mention at least one reason why either model is not ideal.

(f) [*4 marks*] Further extend your final Poisson model to a Negative Binomial model and comment on whether this model is preferable to the other two, on the basis of all the criteria used for comparison so far.