¶ Project

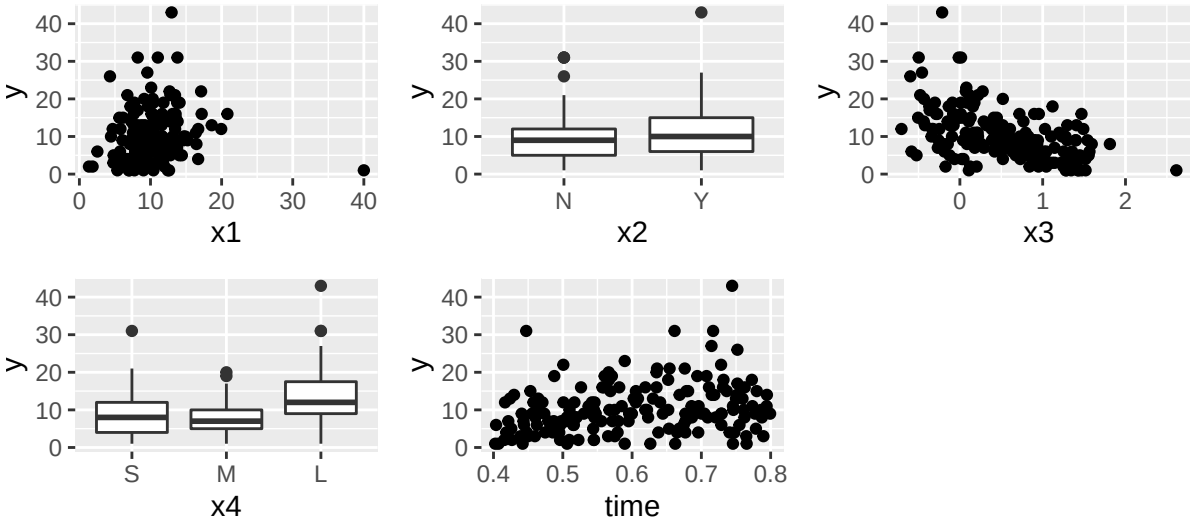
Introduction

The aim of this report is to find a suitable model for the Project2022 dataset given. The response variable is a count of the number of events over a time period, with 4 covariates (x1,x2,x3,x4). The exploration of the data and the statistical methods used will be discussed and all values will be given to 3 significant figures for consistency.

Exploratory Analysis

The first step was to explore and understand the dataset. The dataset had 180 entries and the response variable, y, was strictly positive integers and values typically varied from 1 to 50 with the exception of one value which was 10000. This was an extreme outlier and likely to be the result of a mistake in data creation/handling. It made graphing and further analysis useless as it needed to be removed, therefore entry 107 was deleted from the dataset. Covariates x1 and x3 are continuous variables whereas x2 and x4 are categorical, with 2 and 3 categories respectively. Furthermore, a time period for each each observation is included, where all observations are measured in less than 1 second. No more obvious outliers in entries were observed (except for entry 40 with x1=40 but this is later investigated in the leverage section at the end).

Plots of the covariates vs the response were generated to investigate any visual relationships.



x1 shows a potential positive linear relationship whereas x3 has a possible weak negative linear in comparison. The two categorical variables do not show much of a relationship, only that values "L" for x4 may be slightly higher than others.

Model Fitting

Since the response is strictly positive and represents a count, then the initial model proposed was to be a poisson generalised linear regression model. Since the time period varies for each observation, an offset with respect to time should be included in the model also.

In general, there is some known factor, T_i, which we expect to multiply by the expected count λ_i.

$$Y_i \sim Poisson(\lambda_i T_i),$$

where $log\lambda_i = \eta_i$ as in a regular poisson model, but Ti is the total time over which the count was measured. Notice that

$$logE[Y_i] = log(T_i \exp \eta_i) = \eta_i + logT_i.$$

We, therefore, fit this model using an additive offset where log(time) was taken in our model code.

A forward-fitting method was used, using likelihood ratio tests, to find the best covariates to use in the model. First, a model (M1) with just the intercept and the offset was fitted as a baseline. Next, four models with an intercept, offset and one covariate, x1 through x4 (M1.1-M1.4), were fitted. A likelihood ratio test (LRT) was performed between M1 and each of the single covariate models and the p-values were observed. H0 in this test was that the previous model (M1 in the first case) was the best model and so H0 was rejected if the p-value was less than $\alpha = 0.05$ and the new model with added covariates was taken. The lowest p-value initially was 2.49×10^{-41} for M1.3 so x3 was the most significant covariate and was chosen. We then iterate this process by adding on the covariates x1,x2,x4 to the model M1.3 and performing LRT's again. The smallest p-value was 4.95×10^{-25} for M1.34 so x4 was the next most significant covariate and was chosen. After this, x1 and x2 was tested for and x1 was chosen as it had the smallest p-value of 4.55×10^{-19}. Finally, the full model with all the covariates was tested however the p-value was $0.249 > \alpha$ so we accept H0. This means the poisson model M1.341 with covariates x3, x4 and x1 was chosen to be the best. It is understandable that x2 is not significant in the model due to the nature of the plot earlier showing no real relationship with y. The next step was to look for interactions between the variables. The same forward-fitting method was used as above but this time the models include an interaction on top of the model M1.341. Three models are created testing for all possible interactions between the three covariates x3,x4 and x1. The model with interaction x1:x4 produced the smallest and most significant p-value 2.22×10^{-13} and so this interaction was added to the model. The two remaining interactions were added and tested for against the single interaction model however no p-value was significant therefore H0 was accepted and the poisson model M1.341 with interaction x1:x4 and offset was taken to be the best model.

The AIC and BIC of this model was 962 and 985 respectively, which was the lowest compared with any other model produced so far after exhaustive checking. This shows this version of the poisson model was the best for this data. The deviance however was 242 which is higher than the critical value at the 5% level of 204 (using chi squared distribution) so it was necessary to check other types of generalised linear models that might also be well suited to this data and see if any could produce a lower deviance that would show a better fit. Another appropriate distribution for count data was the negative binomial distribution.

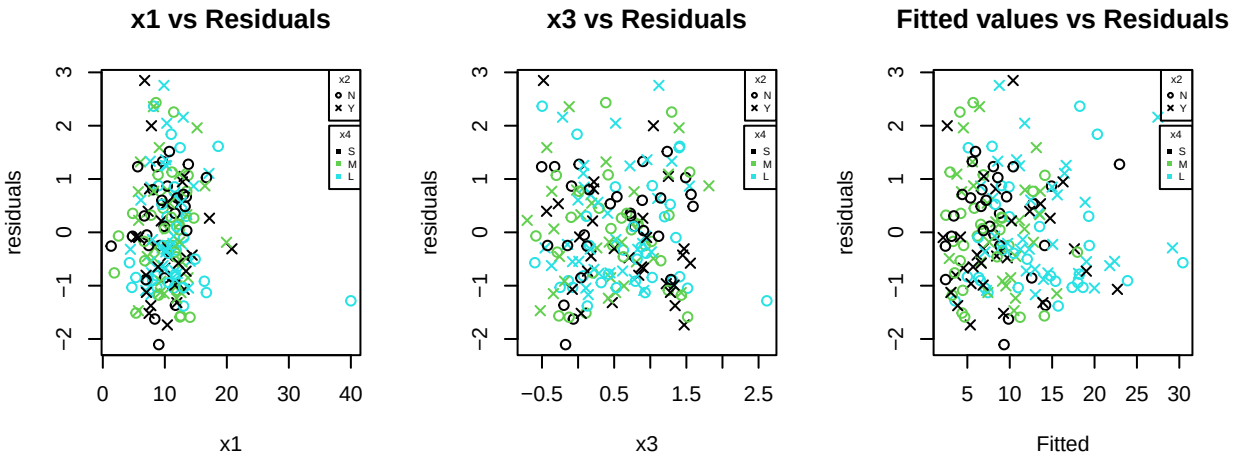
The negative binomial regression is considered a generalisation of poisson regression as it has the same structure however it accounts for additional variance in the mean (poisson assumes mean = variance).[1] We can use negative binomial to estimate a poisson distribution as it approaches large sample size n,

$$Poisson(\lambda) = lim_{r\to\infty} NB(n, \frac{\lambda}{n+\lambda})$$

So NB has larger variance than poisson for small n.[2]

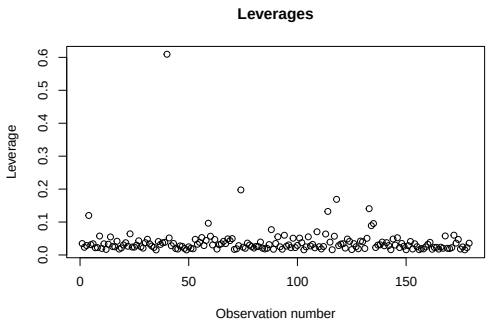
Using the same covariates and interactions from the best poisson model (along with the log offset), the negative binomial equivalent was tested using the AIC and BIC, producing slightly lower values of 955 and 981. More significantly, the deviance produced was 183 which was now much lower and importantly significant showing that this was a better fit. To be confident that this was the best negative binomial model, a backward-fitting method was applied to the covariates. This was the opposite process to forward-fitting; we start with a full model with all the covariates in and then remove each of the covariates seperately and use likelihood ratio tests to see if the models with less covariates was a better fit. Immdeiately, x2 was significant, producing the only p-value $0.388 > \alpha = 0.05$, so x2 was removed. The same process was repeated with the three remaining covariates however all p-values produced were $< \alpha = 0.05$ so the best model definitely included x3, x4 and x1. A forward-fitting method was again applied to find the best covariate interactions. From the first iteration, the lowest p-value was 7.03×10^{-10} corresponding to the interaction x1:x4 and so this was chosen. After the next iteration, it was clear no more interactions are needed due to now significant p-values therefore the negative binomial model we originally looked at was correct. In conclusion, the final model was a **negative binomial with covariates x3, x4, x1 and interaction x1:x4 (as well as accounting for a time offset)**.

After the model was found, various model diagnostics were used to ensure only relevant data was used in model formulation and to assess the quality of the model fit. Firstly, the residuals are plotted against the fitted values from the model, using symbols and colours for the categorical variables.



There was no obvious pattern in the residuals therefore it was likely to be a good fit. After removal of some high residual points (>2), there was no major changes in the covariates with respect to the standard deviations so decision was to keep those rows in despite the high values.

The leverages of each observation are plotted next to see if any values give particular high values The leverage shows how much a point influences the fitting of the model and so high leverage points can cause a worse fit if they are deemed to be potential outliers. Rows with leverages higher than three times the mean leverage are rows 4, 40, 74, 115, 119 and 134. The value at index 40 have a massive leverage of 0.609 so this could be an outlier. From investigation, the x1 value recorded was also 40 so potentially an inputting mistake was made during database creation. This was investigated further by removing all of these rows and finding the model again to see if the coefficients have any noticeable change with respect to their standard deviations. There was no significant changes so the rows were simply kept in the model. (This investigation of leverage was also done at the beginning during the very first model fits, with the same results but the final leverages are only discussed here to avoid repetition)



Results

Finally, the coefficients for the model were (1.48, -0.709, 0.397, 1.71, 0.145, -0.0518, -0.124) for parameters and interactions ((Intercept), x3, x4M, x4L, x1, x4M:x1, x4L:x1). For every unit change in x3, the difference in the logs of expected counts, y, is expected to decrease by 0.709, given that the other covariates are constant. If x4 is M then this increases by 0.397; similarly if L then it increases by 1.71. For every unit change in x1, the log response y increases by 0.145 which was the smallest covariate impact. For each of the interactions, log y only reduces a very small amount of 0.0518 and 0.124 for x4M:x1 and x4L:x1 per unit change respectively so the interactions, as expected, do not have much impact on the model accuracy. These values can all be transformed using exp() to show the effect of covariates on the true y also. The predicted y for (x1, x2, x3, x4) = (10,"N",0.5,"S") was 13.1 with a 95% confidence interval of (11.8,14.6) and the predicted y for (x1, x2, x3, x4) = (10,"N",0.5,"L") was 20.9 with a 95% confidence interval of (19.2,22.7). Values were found by using the model and then transforming, using exp(), the outputs to be proportional to the original data. The big increase in y from changing x4=S to x4=L is a result of x4=L having the biggest covariate impact on y.

1. https://stats.oarc.ucla.edu/r/dae/negative-binomial-regression/
2. https://en.wikipedia.org/wiki/Negative_binomial_distribution#Poisson_distribution