Klaudia Ludwisiak ICA 3 24th March 2017

Introduction

In this coursework we were tasked develop a function to fit a generalised linear model (GLM) for an exponentially distributed response variable using the Iterative Weighted Least Squares approach (IWLS). The exponential function belongs to the exponential family, thus using GLM theory is viable. It should be noted that the exponential distribution is a special case of the gamma distribution where the alpha parameter is equal to one and the beta parameter is equal to lambda. This observation is the foundation fro derivations in the method described below.

The function created is called erm, it's output is in line with the assignment specification and steps undertaken to obtain it are described below. When it comes to data input, the function takes a vector of responses (y) and a design matrix of explanatory variables X. It should be noted that the function does not introduce an intercept term i.e augment the input matrix X with an additional column of 1 values. If the suer wishes to include an intercept term in the model they should augment the matrix X accordingly prior to using the erm function. The last input is the start value for the estimate of the regression coefficient (beta), this is set to zero by default. The start value is the same for all entries in vector beta (length same as y).

Description of steps undertaken in *erm* and the relevant derivations

1) Input checks

(i) Y should be a vector
(ii) X should be a matrix where number of rows (data instances) is the same as length(y) and the number of columns is the number of explanatory variables.
(iii) check (y) data is positive and non-zero. Only this data can be modelled with exponential distribution, which is always positive and asymptotic about the x=0 axis.

2) IWLS implementation for exponentially distributed data

The IWLS implementation was adapted from that specified in lab 8 for a poisson distribution. With some simplifications and using altered distribution parameters, derived from the Gamma distribution with a = 1 and b = λ. The mean was taken at μ*i* = 1/λ*i* and the variance function at V*i* = (1/λ*i*)^2. There was also need to find the derivative of the link function, which was found to be λ. It was then observed that in case of the exponential distribution the expression for the weights gives an identity matrix. i.e.: W*i* = [g′(μ*i*)]^−2 / V*i* = 1. This result allowed for the simplification of the IWLS steps, as in this case, the multiplication of say X by W simply yields X. Further to that the above explained modifications the IWLS procedure was implemented as described in lecture 8.

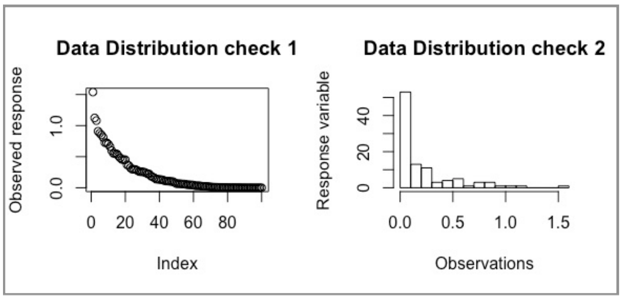
3) Obtaining other required output

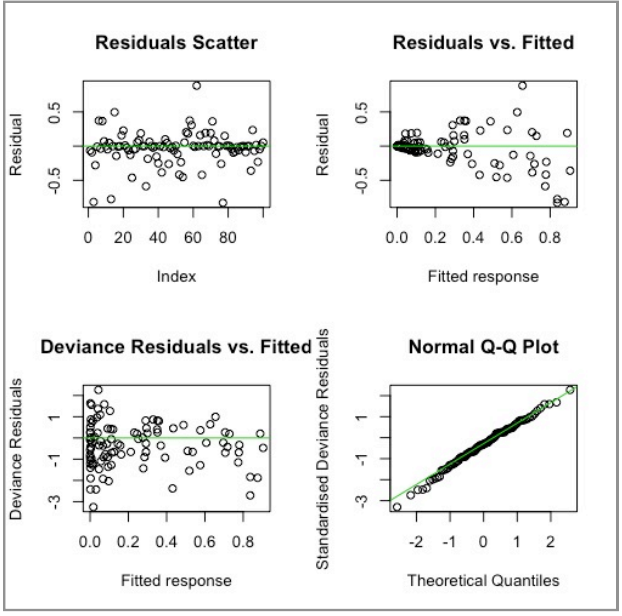
(**i) Variance-covariance matrix** was obtained as a by-product of the implementation of the IWLS procedure, and is equal to the inverse of the information matrix i.e. (X'WX)^(-1)
(ii) Standard error on beta estimates was obtained from the variance- covariance matrix, using the fact that β = sqrt(diag (X'WX))
(iii) Z statistic and P value: Assuming the value of parameter phi is known and equal to 1 we can test using the Z statistic as opposed to the T statistic. The value f the Z statistic is the ratio of the β estimates and their standard error. The resulting P value is found by matching the Z value with the corresponding probability of the normal distribution.
(iv) Fitted values were found by taking the inverse of the exponential of Xβ. These are also equivalent to the expected value of the response variable, hence its mean.
(iiv) Deviance: The fitted values were used to obtain the deviance of the model, where this is defined as : −2 (*l*(*μ*ˆ, λ, y) − *l*(y, λ, y)), where the function *l()* denotes the log-likelihood

and the expression takes the difference of a saturated model with the fitted model. It can be show that for a exponential distribution the expression reduces to:
$D(y, \hat{\mu}) = -2 * SUM * [\ln(y_i/\hat{\mu}_i) - ((y_i - \hat{\mu}_i)/\hat{\mu}_i)],$
where SUM denotes the summation over all *i* data points.
(iiiv) Degrees of Freedom: First it should be noted that the total degrees of freedom of a model are (n-1) where n is the amount of data points. The model degrees of freedom correspond to the number of coefficients estimated minus 1, ie. number of columns in X minus 1. The residual degrees of freedom are the difference between the total degrees of freedom and the model degrees of freedom.

4) Plots and the Calculation of residuals

To check distributional assumptions, two plots are created on same figure; a histogram of the values of y as well as a simple plot of the observed values of y, sorted in a descending order. This is done to enable a visual inspection of the distribution, to check that it indeed resembles the exponential shape. To asses and check the model fit a figure with 4 plots is created. The first plot is a simple residuals scatter where the difference between the 'true' y values and the estimated y values is plotted against the index of the data instance. The closer this difference is to the zero line the better. The second plot is a Residuals vs. Fitted values plot, where the residuals are as defined above. The third plot shows deviance residuals vs fitted values. The fourth plot is the Normal Q-Q plot of standardised deviance residuals vs the theoretical quintiles of the normal distribution.





Testing

The performance of the erm function was evaluated against two datasets, in the following ways:

1) Generating my own exponential data: Here the approach is to first set a value for beta's, then generate a corresponding amount of X variables (with intercept term included as a column of 1's). Using the equation: ln(λ)= Xβ estimate the value of lambda per point. This λi can then be used to sample at random from the exponential distribution and generate the response y, which will have a different lambda for each datapoint.

This approach has the advantage of being able to compare the theoretical β values (assumed when generating dataset) with estimates for β returned by erm. This has been the main way to test the function and the results achieved were satisfactory. For instance, with 100 data points and a theoretical vector beta (0.1, 0.4, 0.3, 0.2) the erm estimates are found at approx. (0.07, 0.39, 0.33, 0.18), upon rounding these match the underlying 'true' β's. Results were further tested using the R glm package, which again verified the correct working of my function.

2) Using an example from datasets: This avenue was attempted but it was tricky to find a dataset which would lend itself to be modelled by an exponential distribution. Thus in the end the data was heavily modified and this test should not be given much weight. Never the less, it proves the working of the my function. Details of this approach can be found in the relevant R code.