Generalised linear models (GLM) in R: Prac5

Question 1.

The impact of climate change on ecological systems world-wide is of growing concern. For example, it is known that increased water temperatures can raise the chances of 'bleaching' events to occur in coral reefs.

Data on the incidence of bleaching events (denoted as 1) in relation to sea water temperature is presented in the file `coral.txt`[1].

Open the prac5Q1.r script file in RStudio and run the commands. You will be running a simple logistic regression model using `bleach` as the response variable and water `temperature` as the explanatory variable. You will

(i) Plot the data.
(ii) Fit a glm and produce the anova table.
(iii) Decide whether the model is a good fit for the data (provide support for your answer).
(iv) Determine whether the regression coefficients are significant (provide support).
(v) Calculate the confidence intervals for the regression coefficients.
(vi) Use predict() to estimate the probability of bleaching when temperature = 30 ^{o}C.
(vii) Verify this result, showing all calculations.
(viii) Superimpose the fitted curve onto the exploratory plot from (a).
(ix) Give a short interpretation of how water temperature relates to whether a coral bleaching event will occur.

Question 2.

You are going to investigate how qualifications and test scores affect whether a given student gets admitted to graduate school. This example is from the University of California, Los Angeles (UCLA) and so the associated university admission credentials are `gre` (Graduate Record Examination – a national standardized test covering verbal, quantitative, and analytic skills to give an overall score between 200 - 800[2]), cumulative undergraduate `gpa` (Grade Point Average – on a scale of 0.0 to 4.0) and `rank` (the relative ranking level of the university from which the student completed their undergraduate degree. Universities were ranked from 1 to 4 with a 1 indicating those with the highest prestige, while those with a rank of 4 having the lowest).

Open prac5Q2.r script file in R and load the online dataset into R. Refer to Q1 for help with the code.

(i) Plot the data for gre and gpa (each as single explanatory factors, i.e., 2 different plots)
(ii) Fit a glm with gre, gpa and rank as predictors and decide if the model is a good fit for the data[3] (provide support for your answer)
(iii) Determine whether the regression coefficients are significant (provide support – and in the case of rank be sure to carefully define what each coefficient refers to *[see below]*)

[1] Adapted from simulated data produced by www.shizukalab.com
[2] Although this scale has now changed to 130 – 170
[3] You might be surprised by the answer … but still continue along!

(iv) Carefully interpret the meaning of all of the regression coefficients

Interpreting regression coefficients for a logistic regression.

Each regression coefficient for rank is a comparison to the baseline level (sometimes called the dummy variable) which, by default, is the first level of the given variable (in this case `rank1` – note that it is omitted from the output). So looking at a portion of your output:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.98998    1.13995   -3.50  0.00047
gre          0.00226    0.00109    2.07  0.03847
gpa          0.80404    0.33182    2.42  0.01539
rank2       -0.67544    0.31649   -2.13  0.03283
rank3       -1.34020    0.34531   -3.88  0.00010
rank4       -1.55146    0.41783   -3.71  0.00020
```

When referring to continuous variables (like `gpa`) we could interpret the regression coefficient by first calculating the exponent (to get the odds):

```
exp(0.80404) = 2.23
```

And then stating: *When gre and rank of university are held constant, the effect of a one unit increase in gpa score increases the odds of being admitted to UCLA by 123% (i.e., because 2.23 – 1 = 1.23 x 100% = 123%). (That sounds good!)*

To interpret categorical variables (like `rank2`) first calculate:

```
exp(-0.675) = 0.509
```

And then state: *When gre and gpa are held constant, the odds of being admitted to UCLA from a rank2 university is approx. 50% less than from a rank1 university (i.e., because 0.509 – 1 = -0.491 x 100% = -49.1%).*

Finally, *only if you've got time* – run the final section of the prac5Q2.r script to use predicted probabilities and graphics to visualize the model results. This code requires the installation of a package into R studio called ggplot2. For instructions on how to load packages check out this link (https://www.youtube.com/watch?v=u1r5XTqrCTQ) or ask me. Also, remember to insert the name of your GLM into the code to ensure that it will run (mine was called "a").

(v) Provide a general interpretation of the four plots of predicted probabilities across differing GPA levels