Generalized Linear Models

STA 310: Homework 5

1. The probability density function (PDF) of the Pareto Type I distribution is given by:

$$f(x) = \frac{\alpha x_m^{\alpha}}{x^{\alpha+1}} \mathbf{1}_{\{x \geq x_m\}}$$

where:

- $\mathbf{1}_{\{x \geq x_m\}}$ is the indicator function, which is 1 if $x \geq x_m$ and 0 otherwise.

- $\alpha > 0$ is the shape parameter.

- $x_m > 0$ is the scale parameter (minimum possible value) of the data.

Is this a member of the exponential family (in canonical form)? If so, identify the components in canonical form. If not, explain why not.

2. Assume a logistic regression model (i.e., binary regression under canonical link). Assume that y_i is binary with success probability p and $i = 1, \ldots, n$. Specifically,

$$y_i \sim \text{Bernoulli}(p_i), \quad \text{where} \quad p_i = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}}. \tag{1}$$

a. Provide the log-likelihood function as a function of the regression parameters.

b. Derive the score functions of the log-likelihood with respect to the regression parameters.

c. Derive the Hessian function of the log-likelihood with respect to the regression parameters.

d. Using the previous steps to write out the Newton-Raphson Algorithm.

e. A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution, effect admission into graduate school. The response variable, admit/don't admit, is a binary variable, which can be read in using the command below.

Consider running a logistic regression using the glm() function in R below. Write your own function to implement the Newton Raphson algorithm to output the regression coefficients, and verify that the coefficients match those from the glm() function.

```
df <- read.csv('https://stats.idre.ucla.edu/stat/data/binary.csv')
mylogit <- glm(admit ~ gre + gpa + rank, data = df, family = "binomial"(link = "logit"))
summary(mylogit)
```

```
##
## Call:
## glm(formula = admit ~ gre + gpa + rank, family = binomial(link = "logit"),
##     data = df)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.449548   1.132846  -3.045  0.00233 **
## gre          0.002294   0.001092   2.101  0.03564 *
## gpa          0.777014   0.327484   2.373  0.01766 *
## rank        -0.560031   0.127137  -4.405 1.06e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 499.98  on 399  degrees of freedom
## Residual deviance: 459.44  on 396  degrees of freedom
## AIC: 467.44
##
## Number of Fisher Scoring iterations: 4
```

```
y<-df[,1]
# add column of 1's for the intercept
x<-as.matrix(cbind(rep(1,400),df[,c(2,3,4)]))

IRLS_logistic <- function(x, y, b.init, tol=1e-8) {
# use the same type of structure for your code
# as we did for the poisson
}
```

3. Propose a homework exercise that you believe would help reinforce the class content to students in this class. You may not repeat any exercise explicitly that was used in class and this exercise should be done individually.

Write and propose the question(s), include a rubric, and include a solution.

You may utilize online resource, but please do cite them if you use them and provide details.