1. **Generalized Linear Model**

   The random variable $Y$ has the Pareto distribution with pdf

   $$f\left(y;\theta\right) = \theta y^{-\theta-1} \tag{1}$$

   (a) Show that this pdf belongs to the exponential family and identify $a(y),\ b(\theta),\ c(\theta),\ d(y)$ as defined in Dobson.

   (b) Is the exponential family of natural exponential family or exponential dispersion family as defined in the lecture?

   (c) Find the score statistic $S(\theta; y)$ and the information $\imath(\theta)$ for (b) above. furthermore, calculate $\mathrm{E}\left[S(\theta;Y)\right]$

   (d) Suppose $Y_1, \ldots, Y_N$ are independent random variables each with the **Pareto** distribution as (1) and

   $$\mathrm{E}\left[Y_i\right] = \left(\beta_0 + \beta_1 x_i\right)^2$$

   i Does this belong to experiential family? Give reasons for your answer.

   ii Is this a generalized Linear Model? Give reasons for your answer.

2. **Logistic Regression** you are given the following information for a GLM of customer retention:

| Response Variable: | retention | |
|---|---|---|
| Response distribution: | Binomial | |
| The link Function: | Logit | |
| Parameter | df | $\beta$ |
| Intercept | 1 | 1.530 |
| Number of Drives | | |
| 1 | 0 | 0 |
| $> 1$ | 1 | 0.735 |
| Last Rate change | | |
| $< 0\%$ | 1 | 0.0160 |
| $[0\%, 10\%]$ | 0 | -0.031 |
| $> 10\%$ | 1 | -0.372 |

   Calculate the probability of retention for a policy with 3 drivers and a prior rate change of 5%.

   Please calculate the value of $\hat{\pi}$.

3. **Logistic Regression** Please refer to Dobson Embryogenic anthers Example 7.4.1 page 133 for the description of example and the data table. The data in Table 1 are the numbers $y_{jk}$ of embryogenic anthers of the plant species obtained when numbers $n_{jk}$ of anthers were prepared under several different conditions. there is qualitative factor with two levels and one continuous explanatory variable by three values.

   The proportions $p_{jk} = y_{jk}/n_{jk}$ are of interested. You are asked to compare three logistic models for $\pi_{jk}$

$$\begin{aligned}\text{Model 1}: \ \text{logit}\ \pi_{jk} &= \alpha_j + \beta_j x_k \\ \text{Model 2}: \ \text{logit}\ \pi_{jk} &= \alpha_j + \beta x_k \\ \text{Model 3}: \ \text{logit}\ \pi_{jk} &= \alpha + \beta x_k\end{aligned}$$

| | | **Centrifuging force (g)** | | |
|---|---|---|---|---|
| **Storage Condition** | | 40 | 150 | 350 |
| Control | $y_{1k}$ | 55 | 52 | 57 |
| | $n_{1k}$ | 102 | 99 | 108 |
| Treatment | $y_{2k}$ | 55 | 50 | 50 |
| | $n_{2k}$ | 76 | 81 | 90 |

Table 1: Embryogenic anther data for question **??**

(a) Run the R to get the coefficients of the three models.

(b) Calculate probability estimates for all three models.

(c) Use the estimated from (b) to calculate the expected values for the three models and match your results with Table 7.7 page 135.

(d) Calculate Pearson residuals using the expected values from (c).

(e) Calculate the following Goodness of fit statistics and comment on their models:

1. $\chi^2$ statistic
2. Deviance D statistic
3. Likelihood Chi-squared statistic C and pseudo $\mathrm{R}^2$

4. **Maximum likelihood estimation approximation** (Iterated weighted least squares Method for estimating parameters) Please refer to Dobson Exercise 4.1 for a description of the data table. We want to estimate parameters using the IWLS method that described in chapter 4 Dobson. The data in Table 2 show the number of cases of AIDS in Australia by date of diagnosis for successive 3-month periods from 1984 to 1998. (Data from National Center for HIV Epidemiology and Clinical Research 1994.)

In this early phase of the epidenic, the numbers of cases seemed to be increasing exponentially.

(a) Plot the number of cases $y_i$, against time period $i$ ($i = 1, 2, \ldots, 20$).

(b) A possible model is the POisson distribution with parameter $\lambda_i = i^\theta$,

$$\log \lambda_i = \theta \log i$$

plot $\log y_i$ against $\log i$ to exam this model.

|  | **Quarter** |  |  |  |
|---|---|---|---|---|
| **Year** | 1 | 2 | 3 | 4 |
| 1984 | 1 | 6 | 16 | 23 |
| 1985 | 27 | 39 | 31 | 30 |
| 1986 | 43 | 51 | 63 | 70 |
| 1987 | 88 | 97 | 91 | 104 |
| 1988 | 110 | 113 | 149 | 159 |

Table 2: Data for question 4

(c) Fit the GLM to these dta using the Poisson distribution, the log-link function and the equation

$$g\left(\lambda_i\right) = \log \lambda_i = \beta_0 + \beta_1 x_i,$$

where $x_i = \log i$. Firstly, do this from first principles, working out expressions for the weighted matrix $\mathbf{W}$ and other terms needed for the iterative equation

$$\mathbf{X}^T \mathbf{W} \mathbf{X} \mathbf{b}^m = \mathbf{X}^T \mathbf{W} \mathbf{z}$$

and using software which can perform matrix operations to carry out the calculation. please show the first 3 steps of your work on the following:

1. $\mathbf{b^m}$, for example, the initial step, $b_0^0, b_1^0$
2. Score function $\mathbf{U}^m$ and information matrix $\imath^m$
3. Weighted matrix $\mathbf{W}$ and $\mathbf{z}$ at step $m$

5. **Deviance** You are given the following information about a GLM:

- The response variables $y_1, y_2, \ldots, y_n$ follow independent exponential distributions with unknown means $\mu_1, \mu_2, \ldots, \mu_n$.
- The fitted means $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_n$

**Calculation of Pearson and Deviance residuals** A set of observations, $y_1, y_2, \ldots, y_n$, are assumed to be exponentially distributed,

$$f(y_i) = \frac{e^{-y_i/\theta_i}}{\theta_i}, i = 1, 2, \ldots, n$$

A GLM is fit to the data with the following model specification:

$$\ln(\theta_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i2}^2 + \beta_4 x_{i3}$$

The vector of observed response, $\mathbf{y}$ and the design matrix, $\mathbf{x}$, are given for the first four observations as well as the vector of all estimated parameters, $\hat{\beta}$:

$$\mathbf{y} = \begin{pmatrix} 15 \\ 85 \\ 10 \\ 40 \end{pmatrix}, \mathbf{x} = \begin{pmatrix} 1 & 1 & 2.5 & 6.25 & 1 \\ 1 & 0 & 1.5 & 2.25 & 1 \\ 1 & 1 & 2.9 & 8.41 & 1 \\ 1 & 1 & 3.0 & 9.00 & 0 \end{pmatrix}, \hat{\beta} = \begin{pmatrix} 2.99 \\ -0.27 \\ -0.67 \\ 0.16 \\ 0.91 \end{pmatrix}$$

(a) What is the deviance, $D$?

(b) Calculate the Pearson residual for the second observation, $r_2$.

(c) Calculate the deviance residual for the second observation, $d_2$.

6. **Nominal Regression**

A nominal logistic model is used to predict the type of car one buys. The base category of car is "sedan". The other categories are "van" and "SUV". The explanatory variables are "gender" and "age group". The fitted coefficients are:

| Type of Car | Van | SUV |
|---|---|---|
| Intercept | 0.10 | -0.02 |
| Gender | | |
| Male | 0 | 0 |
| Female | -0.18 | -0.06 |
| Age | | |
| Under 25 | -0.11 | 0.18 |
| 25-54 | 0 | 0 |
| 45 and up | 0.06 | 0.04 |

(a) (5 points) Calculate the odds ratio of females for vans.

(b) (5 points) Calculate the probability a male age 20 buys an SUV.

7. **Ordinal Regression**

An ordinal variable classifies drivers as follows:

| j | Category |
|---|---|
| 1 | Low Risk |
| 2 | Medium Risk |
| 3 | High Risk |

A cumulative logit model is constructed and the following are the fitted coefficients:

| Category | Low Risk | Medium Risk |
|---|---|---|
| Intercept | 1.30 | 2.05 |
| Gender | | |
| Male | 0 | 0 |
| Female | 0.75 | 0.80 |
| Age | | |
| Under 25 | -1.80 | -0.65 |
| 25-44 | 0 | 0 |
| 45-64 | 1.00 | 1.47 |
| 65 and over | 0.23 | -0.12 |

Calculate the probability that a male driver age 65 or over is a medium risk.