1. Iteratively Re-weighted Least-Squares and Logistic Regression

For this problem, we will employ a data set from a Dutch study of lung cancer from 1985. The data set is in the file “LungCancerBirdKeeping.Rdata” on the class webpage.

This is a case-control study looking at various potential effects on the probability of contracting lung cancer. There are 147 records with variables

- LC: indicator of lung cancer status (yes = 1)
- SEX: patients gender (1 = female)
- SS: indicator of high socio-economic status
- BK: indicator of whether subject kept a pet bird
- AG: subject’s current age
- YR: number of years the subject smoked cigarettes
- CD: average number of cigarettes smoked per day.

We will treat LC as the outcome of interest.

The surprising conclusions of the study (though we won’t investigate this here) is that bird keeping comes out as being a strong risk factor for lung cancer.

(a) Write a program to carry out iteratively re-weighted least squares for logistic regression using only the `lm` function in R. Report your code; including details of initial starting values and stopping criteria.

(b) Use your code to estimate parameters for the data set. Using the output of your final iteration, conduct a Wald test to determine whether the coefficients (other than the intercept) are collectively different from zero.

(c) Evaluate the likelihood and its derivative with the coefficients (other than the intercept) set to zero. Obtain likelihood ratio and score tests and compare these to the Wald test above.

(d) Re-run your estimation procedure with starting values of zero for all coefficients (including the intercept). Record parameter values at each iteration of IWRLS and plot the rate at which they converge – does this appear to be approximately quadratic?

(e) Convert your code to use the complementary log-log link instead of a logistic link and repeat part 1d above.

(f) Given that this is a retrospective study (the study selected patients under age 65 with lung cancer in The Hague, and also selected 98 controls from the town with the same general age structure), what assumptions are needed to make the findings of this study valid? Can they be assessed from the data?
2. Miscellaneous Geometry of Likelihood

(a) **Reparameterization**: we consider transformations of the coefficients $\gamma = G(\beta)$ for $G : \mathbb{R}^{p+1} \to \mathbb{R}^{p+1}$ an invertible function. Demonstrate that the likelihood ratio and score statistics are invariant to this reparameterization; for which re-parameterizations is the Wald test also invariant?

(b) **Convexity and convergence guarantees**: derive conditions under which the score function (ie the derivative of the log likelihood) is convex in a generalized linear model and is strictly monotone over straight lines in coefficient space (hint: this requires conditions on $b(\theta(\eta))$).

Hence demonstrate that under these conditions IWRLS with a canonical link converges to its optimum from any starting values. (It may be easiest to think about this for a one-dimensional optimization first.)

3. Gamma GLMs

This question considers the use of the Gamma distribution in generalized linear models. Recall that the Gamma distribution has density

$$f(y; \beta, \lambda) = \frac{y^{\lambda-1}e^{-y/\beta}}{\Gamma(\lambda)\beta^\lambda}, \quad y > 0$$

for shape parameter $\lambda > 0$ and scale parameter $\beta > 0$. The fact Gamma random variables are always positive makes this distribution attractive for modeling positive, continuous-valued observations.

In particular, at $\lambda = 1$ we recover the exponential distribution, making the Gamma distribution a more flexible family.

(a) Gamma and Exponential Families

i. Write out the Gamma density in exponential family form as used in class and McCullagh and Nelder, what is the appropriate reparameterization of $(\lambda, \beta)$ in order to do this?

ii. Derive the cumulant generator $b(\theta)$, mean and variance along with the canonical link for a Gamma generalized linear model.

iii. Does the range of the mean correspond to the domain of the canonical link if it is taken over an unrestricted domain? Suggest an alternative link to ensure that the GLM does not violate restrictions on the range of Gamma parameters.

(b) Distribution of Gamma residuals.

i. Derive explicit expressions for the Pearson and Deviance residuals for a Gamma GLM.

ii. Demonstrate that that the distribution of $Y/\mu$ does not depend on $\mu$ and further that the distribution of both Pearson and Deviance residuals also does not depend on $\mu$.

iii. Plot the density of the Pearson and Deviance residuals for the values of $\phi$ given by 0.1, 0.5, 1, 2, 5 and 10. Hence evaluate the quality of their normal approximation. (Note: if you know the density of $X$ is $f(x)$, the density of $s(X)$ is $s'(x)f(s(x))$ for a monotone function $s$.)
iv. Comment on the suitability of Pearson versus Deviance residuals for assessing lack of fit; both visually and via a likelihood ratio test (if φ is known).

(c) Estimating φ: The log likelihood of φ does not admit a closed-form expression for its MLE. Two alternative estimates are the mean square of the Pearson residuals and the mean square of the Deviance residuals.

Demonstrate that while the mean square of the Pearson residuals is a consistent estimate of φ, the mean square of the Deviance residuals is biased. Plot this bias; it will be helpful to know that $E \log(Y) = \psi(1/\phi) - \log(1/\phi)$ where $\psi$ is the digamma function.