MSCA31010: Linear & Non-Linear Models

Winter 2021 Assignment 1

Question 1 (40 points)

Dr. Maurice Tweedie introduced the Inverse Gaussian distribution in his 1945 paper in the British weekly scientific journal *Nature*. He later discussed this distribution in great detail in his 1956 paper titled "Statistical Properties of Inverse Gaussian Distributions. I". Although the name contains the word Gaussian, the Inverse Gaussian distribution is unrelated to the Gaussian distribution.

The Inverse Gaussian distribution has two parameters, namely, $\lambda > 0$ and $\mu > 0$. The probability density function (see below) of the Inverse Gaussian distribution is defined for $y > 0$.

$$f(y) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left(-\frac{\lambda(y-\mu)^2}{2y\mu^2}\right)$$

a) (20 points). Please express the probability density function in the form

$$f(y) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}$$

for the Exponential family of distribution. What are θ, ϕ, $a(\phi)$, $b(\theta)$, and $c(y, \phi)$?

b) (10 points). What is the canonical link function for the Inverse Gaussian distribution?

c) (10 points). Please search on the Internet to find one recent application for the Inverse Gaussian distribution. You need to provide a brief description and the reference link of the application.

Question 2 (20 points)

Consider a generalized linear model with the Poisson distribution and the canonical link function (i.e., logarithm). The log-likelihood function with respect to $\boldsymbol{\beta}$

$$l(\boldsymbol{\beta}, \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{n} \left(y_i(\mathbf{x}_i^{\mathrm{t}}\boldsymbol{\beta}) - \exp(\mathbf{x}_i^{\mathrm{t}}\boldsymbol{\beta}) - \ln(y_i!) \right)$$

Suppose there is only $p = 1$ element in the column vectors $\mathbf{x}_i$ and $\boldsymbol{\beta}$. Furthermore, assume that $x_{ij} = 1$. In another word, we do not use any predictors to predict the target variable. The only term in the model is the Intercept term.

a) (10 points). The maximum likelihood estimate (MLE) for β_1 can be found explicitly. Please provide the algebraic expression of the MLE.

b) (10 points). What is the maximum value of the likelihood function under this MLE?

Question 3 (40 points)

Train a generalized linear model on the claim_history.csv where CAR_USE is 'Private'. The target variable is CLM_COUNT. The continuous predictors are: AGE, CAR_AGE, HOMEKIDS, KIDSDRIV, MVR_PTS, TIF, TRAVTIME, and YOJ. The distribution is Poisson and the canonical logarithm link function. You need to use perform the iterations using your own Python codes. You need to drop all missing values (i.e., NaN) of all the predictors and the target variable before training your model.

a) (5 points). How many complete cases did you retain for training your model?

b) (10 points). You can iterate as many times as you need. After the iteration has converged, then please present your Iteration History Table.

c) (10 points). What are the parameter estimates, the asymptotic standard errors, and the 95% asymptotic confidence intervals?

d) (10 points). What is the asymptotic correlation matrix?

e) (5 points). Please comment on the asymptotic correlation matrix.