Exercise 15

a) Density of the Inverse Gaussian distribution can be transformed as:

$$f(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi y^3}\sigma}\exp\left(-\frac{1}{2y}\left(\frac{y-\mu}{\mu\sigma}\right)^2\right) = \left[\frac{1}{\sqrt{2\pi y^3}\sigma} =: k(y,\sigma)\right] =$$

$$= k(y,\sigma)\exp\left(-\frac{y^2-2\mu y+\mu^2}{2y\mu^2\sigma^2}\right) = k(y,\sigma)\exp\left(-\frac{1}{2}\left(\frac{y}{\mu^2\sigma^2}-\frac{2}{\mu\sigma^2}+\frac{1}{y\sigma^2}\right)\right) =$$

$$= \left[c(y,\sigma) := k(y,\sigma)\exp\left(-\frac{1}{2}\left(\frac{1}{y\sigma^2}\right)\right)\right] = c(y,\sigma)\exp\left(\frac{-\frac{y}{2\mu^2}+\frac{1}{\mu}}{\sigma^2}\right) =$$

$$= \left[\phi=\sigma^2, \theta = -\frac{1}{2\mu^2}, a(\theta) = -\sqrt{-2\theta} = \frac{1}{\mu}\right] = c(y,\phi)\exp\left(\frac{\theta y - a(\theta)}{\phi}\right)$$

Thus Inverse Gaussian distribution with $[\mu,\sigma^2]$ parametrization belongs to the exponential family with $\phi=\sigma^2, \theta = -\frac{1}{2\mu^2}, a(\theta) = -\sqrt{-2\theta}$.

b) We remember that for the Exponential Family distributions mean is $E[y] = \mu = \dot{a}(\theta)$ and variance function is $V[\mu] = \ddot{a}(\theta)$, which are thus in our case equal to $\mu = \dot{a}(\theta) = \frac{2}{2\sqrt{-2\theta}} = \frac{1}{\sqrt{-2\theta}}$ and $V[\mu] = \ddot{a}(\theta) = \frac{-0.5\times(-2)}{\sqrt{(-2\theta)^3}} = \frac{1}{\sqrt{(-2\theta)^3}} = \mu^3$ correspondingly.

c) The canonical link has the property that $g(\mu) = \theta = \eta = x^T\beta$ it is invariant to being multiplied by the constant in terms of remaining canonical, i.e. $g^*(\mu) = kg(\mu) = kx^T\beta = x^T\gamma$ is also canonical. Since in our case we have shown that $\theta = -\frac{1}{2\mu^2}$ then $g(\mu) = -\frac{1}{2\mu^2}$ or alternatively $g^*(\mu) = \frac{1}{\mu^2}$ is also a canonical link for the Inverse Gaussian responses.

Exercise 14

As a reminder to the lectures and addition to the discussion we had on plenaries.

We will in this exercise look at the connection between *weighted least squares* and the *Fisher-scoring algorithm* for GLMs (the iterative algorithm used to estimate the MLE of β).

First look at the *weighted least squares*:

a) Consider the linear model (identity link):

$$y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i, \qquad i = 1,\ldots,n$$

The same regression model written on matrix form is:

$$\boldsymbol{Y}_{n*1} = \boldsymbol{X}_{n*(p+1)}\boldsymbol{\beta}_{(p+1)*1} + \boldsymbol{\epsilon}_{n*1}$$

Here $x_{i0} = 1$ and $\epsilon_i \sim N(0,\sigma^2/w_i)$, where w_i signals the precision in observation i.

Let

$$y_i^* = \sqrt{w_i}y_i, \qquad x_{ij}^* = \sqrt{w_i}x_{ij}, \qquad \varepsilon_i^* = \sqrt{w_i}\varepsilon_i,, \qquad i = 1,\ldots,n, \qquad j = 0,\ldots,p$$

Show that we now can write a regression model with y_i^* as response and x_{ij}^* as covariates where

the noise terms have constant variance:

$$\sqrt{w_i} \cdot y_i = \sqrt{w_i} \cdot \left(\beta_0 x_{i0} + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i \right)$$
$$y_i^* = \beta_0 x_{ij}^* + \cdots + \beta_p x_{ip}^* + \varepsilon_i^*$$

where

$$E[\varepsilon_i^*] = \sqrt{w_i} \cdot E[\varepsilon_i] = 0 \qquad \text{Var}[\varepsilon_i^*] = \text{Var}[\sqrt{w_i}\varepsilon_i] = w_i \text{Var}[\varepsilon_i] = \sigma^2.$$

So that $\boldsymbol{Y}^* \sim N(\boldsymbol{X}^*\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$.

b) Calculate the least squares estimate $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$:
Begin with the formulation using $\boldsymbol{Y}^*$

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^{*T}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*T}\boldsymbol{Y}^* \\ &= ((\sqrt{\boldsymbol{W}}\boldsymbol{X})^T\sqrt{\boldsymbol{W}}\boldsymbol{X})^{-1}(\sqrt{\boldsymbol{W}}\boldsymbol{X})^T\sqrt{\boldsymbol{W}}\boldsymbol{Y} \\ &= (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Y}\end{aligned}$$

where $\boldsymbol{W} = diag\{w_i\}$.

We will now turn to GLMs:
c) We remember from class (page 67 in J and H) that the Score function and Expected Fisher info. for GLM are:

$$s_j(\boldsymbol{\beta}) = \frac{1}{\phi}\sum_{i=1}^{n} x_{ij}\frac{y_i - \mu_i}{g'(\mu_i)V(\mu_i)}, \quad I_{j,k}(\boldsymbol{\beta}) = \frac{1}{\phi}\sum_{i=1}^{n}\frac{x_{ij}x_{ik}}{g'(\mu_i)^2 V(\mu_i)}$$

where the μ_i's are indirectly specified through $\boldsymbol{\beta}$. $V = \frac{d^2 a(\theta_i)}{(d\theta_i)^2}$

Writing it on matrix form: $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\beta}) = (\mu_1, \ldots, \mu_p)^T, \boldsymbol{G}(\boldsymbol{\beta}) = diag\{g'(\mu_i)\}, \boldsymbol{W}(\boldsymbol{\beta}) = diag\{\frac{1}{g'(\mu_i)^2 V(\mu_i)}\}$

Show that we now can express s and I in terms of:

$$\begin{aligned}\boldsymbol{s}(\boldsymbol{\beta}) =& \frac{1}{\phi}\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta})\boldsymbol{G}(\boldsymbol{\beta})(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta})) \\ =& \frac{1}{\phi}\boldsymbol{X}^T diag\left(\frac{1}{g'(\mu_i)^2 V(\mu_i)}\right) diag\left(g'(\mu_i)\right)(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta})) \\ =& \frac{1}{\phi}\boldsymbol{X}^T diag\left(\frac{1}{g'(\mu_i)V(\mu_i)}\right)(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta})) \\ =& \frac{1}{\phi}\begin{bmatrix} x_{10}\frac{1}{g'(\mu_1)V(\mu_1)} & \cdots & x_{n0}\frac{1}{g'(\mu_n)V(\mu_n)} \\ \vdots & & \vdots \\ x_{1p}\frac{1}{g'(\mu_1)V(\mu_1)} & \cdots & x_{np}\frac{1}{g'(\mu_n)V(\mu_n)} \end{bmatrix}\begin{bmatrix} y_i - \mu_i \\ \vdots \\ y_n - \mu_n \end{bmatrix} \\ =& \frac{1}{\phi}\begin{bmatrix} \sum_{i=1}^n x_{i0}\frac{y_i - \mu_i}{g'(\mu_i)V(\mu_i)} \\ \vdots \\ \sum_{i=1}^n x_{ip}\frac{y_i - \mu_i}{g'(\mu_i)V(\mu_i)} \end{bmatrix}\end{aligned}$$

Hence,

$$s_j(\boldsymbol{\beta}) = \frac{1}{\phi}\sum_{i=1}^{n} x_{ij}\frac{y_i-\mu_i}{g'(\mu_i)V(\mu_i)} \qquad j=0,\dots,p.$$

$$\begin{aligned}
\boldsymbol{I}(\boldsymbol{\beta}) = \frac{1}{\phi}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} &= \frac{1}{\phi}\boldsymbol{X}^T \begin{bmatrix} \frac{x_{1,0}}{g'(\mu_1)^2V(\mu_1)} & \cdots & \frac{x_{1,(p+1)}}{g'(\mu_1)^2V(\mu_1)} \\ \vdots & & \vdots \\ \frac{x_{n,0}}{g'(\mu_n)^2V(\mu_n)} & \cdots & \frac{x_{n,(p+1)}}{g'(\mu_n)^2V(\mu_n)} \end{bmatrix} \\
&= \frac{1}{\phi}\begin{bmatrix} \sum_{i=1}^n \frac{x_{i,0}x_{i,0}}{g'(\mu_i)^2V(\mu_i)} & \cdots & \sum_{i=1}^n \frac{x_{i,0}x_{i,(p+1)}}{g'(\mu_i)^2V(\mu_i)} \\ \vdots & \sum_{i=1}^n \frac{x_{ij}x_{ik}}{g'(\mu_i)^2V(\mu_i)} & \vdots \\ \sum_{i=1}^n \frac{x_{i,(p+1)}x_{i,0}}{g'(\mu_i)^2V(\mu_i)} & \cdots & \sum_{i=1}^n \frac{x_{i,(p+1)}x_{i,(p+1)}}{g'(\mu_i)^2V(\mu_i)} \end{bmatrix} \\
I_{jk} &= \frac{1}{\phi}\{\sum_{i=1}^{n}\frac{x_{ij}x_{ik}}{g'(\mu_i)^2V(\mu_i)}\}
\end{aligned}$$

(It says J in the exercise, but it should be I. J is often used as the observed fisher info)

d) Now we want to find an estimate of β.
Show that the Fisher scoring algorithm can be written as:

$$\boldsymbol{\beta}^{(k+1)} = (\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta}^{(k)})\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta}^{(k)})\boldsymbol{z}^{(k)}$$

where

$$\boldsymbol{z}^{(k)} = \boldsymbol{X}\boldsymbol{\beta}^{(k)} + \boldsymbol{G}(\boldsymbol{\beta}^{(k)})\big(\boldsymbol{y}-\boldsymbol{\mu}(\boldsymbol{\beta}^{(k)})\big)$$

Begin with the Newton-Rahpson iteration: (page 69)

$$\begin{aligned}
\boldsymbol{\beta}^{(k+1)} &= \boldsymbol{\beta}^{(k)} - l''(\boldsymbol{\beta}^{(k)})^{-1}l'(\boldsymbol{\beta}^{(k)}) \\
&= \boldsymbol{\beta}^{(k)} + \boldsymbol{J}(\boldsymbol{\beta}^{(k)})^{-1}\boldsymbol{s}(\boldsymbol{\beta}^{(k)})
\end{aligned}$$

Fisher suggested to replace $\boldsymbol{J}$ by its expectation, thus $\boldsymbol{I}$. Using the previous results for $\boldsymbol{s}$ and $\boldsymbol{I}$ we get:

$$\begin{aligned}
\boldsymbol{\beta}^{(k+1)} &= \boldsymbol{\beta}^{(k)} + \left(\frac{1}{\phi}\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta}^{(k)})\boldsymbol{X}\right)^{-1}\frac{1}{\phi}\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta}^{(k)})\boldsymbol{G}(\boldsymbol{\beta}^{(k)})\left(\boldsymbol{y}-\boldsymbol{\mu}(\boldsymbol{\beta}^{(k)})\right) \\
&= \boldsymbol{\beta}^{(k)} + \left(\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta}^{(k)})\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta}^{(k)})[\boldsymbol{G}(\boldsymbol{\beta}^{(k)})(\boldsymbol{y}-\boldsymbol{\mu}(\boldsymbol{\beta}^{(k)})) + \boldsymbol{X}\boldsymbol{\beta}^{(k)} - \boldsymbol{X}\boldsymbol{\beta}^{(k)}] \\
&= \boldsymbol{\beta}^{(k)} + \left(\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta}^{(k)})\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta}^{(k)})[\boldsymbol{z}^{(k)} - \boldsymbol{X}\boldsymbol{\beta}^{(k)}]
\end{aligned}$$

Let $\boldsymbol{A} = \boldsymbol{X}^T \boldsymbol{W}(\boldsymbol{\beta}^{(k)})\boldsymbol{X}$

$$\begin{aligned}
\boldsymbol{\beta}^{(k+1)} &= \boldsymbol{\beta}^{(k)} + \boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta}^{(k)})[\boldsymbol{z}^{(k)} - \boldsymbol{X}\boldsymbol{\beta}^{(k)}] \\
&= \boldsymbol{\beta}^{(k)} + \boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta}^{(k)})\boldsymbol{z}^{(k)} - \boldsymbol{A}^{-1}\boldsymbol{A}\beta^{(k)} \\
&= \boldsymbol{\beta}^{(k)} + \boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta}^{(k)})\boldsymbol{z}^{(k)} - \beta^{(k)} \\
&= \boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta}^{(k)})\boldsymbol{z}^{(k)} \\
&= (\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta}^{(k)})\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta}^{(k)})\boldsymbol{z}^{(k)}
\end{aligned}$$

To find the estimate of $\boldsymbol{\beta}$ in a GLM setting, we start with a quess $\boldsymbol{\beta}^{(0)}$, calculate $\boldsymbol{W}(\boldsymbol{\beta}^{(0)})$ and $\boldsymbol{z}^{(0)}$, calculate a new guess $\boldsymbol{\beta}^{(1)}$ and so on ... until we only see a small change in the update of $\boldsymbol{\beta}$.

The expression from the Fisher algorithm is similar to the weighted least square estimation. For simplicity we show the following results for one observation and for β being a scalar.
Looking at a link function $g(\mu) = x\beta$ we find, by Taylor expansion:

$$\begin{aligned}
g(y) &\approx g(\mu) + g'(\mu)(y-\mu) \\
g'(\mu) &\approx \frac{g(y) - g(\mu)}{y - \mu} \\
z = x\beta + g'(\mu)(y-\mu) &\approx g(\mu) + \frac{g(y) - g(\mu)}{y-\mu}(y-\mu) = g(y)
\end{aligned}$$

So $\boldsymbol{\beta}^{(k+1)} \approx (\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta}^{(k)})\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta}^{(k)})g(\boldsymbol{Y})$.

For the identity link: $g(y) = y$ and $g'(\mu) = 1$ we get the weighted least square formula.

$$z = x\beta + 1(y-\mu) = \mu + y - \mu = y$$

e) Looking at an example:
Assume now that $Y_i \sim N(\mu_i, \sigma^2)$ with $\mu_i = \sum_{j=0}^{p} \beta_j x_{ij}$.
Show that the Fisher scoring algorithm converges in the first iteration to the least squares estimator.

The expectation $E(y_i) = \mu_i = \sum_{j=0}^{p} \beta_j x_{ij} = \eta_i$, thus we have the identity link: $g(\mu_i) = \mu_i$.

From the normal distribution (we have showed earlier that)

$$\theta_i = \mu_i, \qquad a(\theta_i) = \frac{1}{2}\mu_i^2 = \frac{1}{2}\theta_i^2, \qquad a'(\theta_i) = \theta_i, \qquad a''(\theta) = 1,$$

and

$$V(\mu) = a''(\theta) = 1,$$

such that

$$\boldsymbol{W} = diag\{\frac{1}{g'(\mu_i)^2 V(\mu_i)}\} = diag(1)$$

Start with $\boldsymbol{\beta}^{(0)}, \quad \boldsymbol{\mu}(\boldsymbol{\beta}^{(0)}) = \boldsymbol{X}\boldsymbol{\beta}^{(0)}$

$$\begin{aligned}
\boldsymbol{\beta}^{(1)} &= (\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta}^{(0)})\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}(\beta^{(0)})\boldsymbol{z}(\boldsymbol{\beta}^{(0)}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{z}(\boldsymbol{\beta}^{(0)}) \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\Big(\boldsymbol{X}\boldsymbol{\beta}^{(0)} + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(0)})\Big) \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}
\end{aligned}$$

which is the least square solution!