2014, Q5 b)

Q For each situation below, state whether the model can be regarded as a GLM - give reasons.

Maths (2): i) $Y_i \in \{0,1\}$, $\mathbb{P}(Y_i = 1) = e^{-\beta x_i}$

(2) ii) $Y_i \sim$ exponential, with mean μ_i, $\mu_i = e^{\beta_0 + \beta_1 x_i}$

(2) iii) $Y_i \sim \text{Poisson}(\mu_i)$, $\mu_i = \alpha \cdot e^{\beta x_i}$, ($\alpha$ a known constant > 0)

(2) iv) $Y_i \sim N(\mu_i, \sigma^2)$ with $\mu_i = \alpha + \log(\beta_0 + \beta_1 x_i)$, with α unknown. ← This is $\mu_i = \mathbb{E}(Y_i)$.

Solution:

i) Yes since Bernoulli is a member of exponential family and $g(\mathbb{E}(Y_i)) = g(\mathbb{P}(Y_i = 1)) = -\beta x_i$

$\equiv \beta x_i$, which is linear.

where $g(.)$ is the log link.

ii) Yes, the exponential is a member of the exponential family and $\log(\mu_i) = \beta_0 + \beta_1 x_i$

$\equiv$ a linear predictor.

iii) Yes, Poisson $\in$ {expo-family} and again use log link: $g(\mu_i) = \eta_i =$

$\log(\mu_i) + \beta x_i = \text{const} + \beta x_i$

iv) NO, although Normal is a member of expo-family, there is no way to do $g(\alpha + \log(\beta_0 + \beta_1 x_i)) = (\alpha, \beta_0, \beta_1)\begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix}$ + const

where g is a valid link.

2014, Q8 b)

- Let $X_1, X_2, \dots, X_6 \overset{i.i.d}{\sim} N(\mu_1, \Sigma)$, $X_i \in \mathbb{R}^2$, $\mu_1 \in \mathbb{R}^2$, $\Sigma \in \mathbb{R}^{2\times 2}$
- Let $Z_1, Z_2, \dots, Z_7 \overset{i.i.d}{\sim} N(\mu_2, \Sigma)$ with $Z_i \in \mathbb{R}^2$, $\mu_2 \in \mathbb{R}^2$
- $\{X_1, \dots, X_6\} \perp\!\!\!\perp \{Z_1, \dots, Z_7\}$
- Sample statistics
  - $\hat{\mu}_1 = \begin{pmatrix} 6 \\ 3 \end{pmatrix}$, $\hat{\mu}_2 = \begin{pmatrix} 2 \\ 9 \end{pmatrix}$, MLE estimators
  - $S_u = \begin{pmatrix} 3 & -2 \\ -2 & 4 \end{pmatrix}$ pooled, unbiased estimator of Σ.
- Let $T^2 = \frac{6 \cdot 7}{6+7} (\hat{\mu}_1 - \hat{\mu}_2)^T S_u^{-1} (\hat{\mu}_1 - \hat{\mu}_2)$.
- Q: use T^2 to test $H_0 : \mu_1 = \mu_2$. (6 marks)
- Ans: $\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$ $\Rightarrow T^2 = 30.69$
  - Under H_0, $T^2 \sim T_p^2(6+7-2)$
  - $T_p^2(n) \equiv \frac{np}{n-p+1} F(p, n-p+1)$
  - $\therefore F \equiv \frac{11-2+1}{11 \cdot 2} T^2 = 13.95 \overset{H_0}{\sim} F(2, 10)$

From tables, $0.05 = \alpha = \mathbb{P}_{H_0}(\text{error}) = \mathbb{P}_{H_0}(F > k)$

entails setting $k = 4.1$ $\therefore$ $\mu_1 \neq \mu_2$ at the $1-0.05$ level of significance.

2014, Q8 c) 6 marks

Let $f_1(x)$, $f_2(x)$ ~~are~~ be two multivariate normal densities: $N_p(\mu_1, \Sigma)$ and $N_p(\mu_2, \Sigma)$ respectively. Same as in part b).

- Consider the classifier:

Allocate x to pop 1 iff

$$l^T x - \frac{1}{2} l^T (\mu_1 + \mu_2) \geq k,$$

where $l = \Sigma^{-1}(\mu_1 - \mu_2)$ and $k = \log \frac{\pi_2}{\pi_1} \cdot \frac{C_{1|2}}{C_{2|1}}$

- π_i = prior probability x is from pop i,
- $C_{i|j}$ = the misclassification cost of saying i when the truth is j.

Q: using the results from part b), if

$x = \begin{pmatrix} 4 \\ 7 \end{pmatrix}$ and $\pi_1 = 0.7$ and $C_{1|2} = C_{2|1}$,

classify x.

A: Plug in $\hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma} = S_u$,

- $\hat{l}^T x - \frac{1}{2} \hat{l}^T (\hat{\mu}_1 + \hat{\mu}_2) = -1.25$
- $k = \log 7/3 = 0.85$

$\therefore$ Allocate x to pop 2.

GLM - Example - SOLUTIONS -

a) Let Y_i be the i^{th} response. $\log \frac{\pi_i}{1-\pi_i} = \eta_i = \beta^T x_i$ under logistic regression, where x_i encodes both age and breathing status. The plot shows $\log \tilde{\pi}_i / 1-\tilde{\pi}_i$

plotted against log(age) and breath, where

$\tilde{\pi}_i = \frac{\text{wheezi}}{\text{total } i}$. From the plots, a logistic regression could work if it is of the form:

$$\eta_i = \begin{cases} \beta_0 + \beta_1 \log(age_i) & \text{, if breath}_i = B \\ \beta_0 + \beta_2 + (\beta_1 + \beta_3)\log(age_i) & \text{, if breath}_i = N \end{cases}$$

i.e. the plots show a clear need for a different intercept depending on breathing status and it is also clear that an interaction between log(age) and breath is needed.

b). Model 2 is what is fitted above, i.e,

$$Y_i \sim Bin(n_i, \pi_i), \quad \log \pi_i / 1-\pi_i = \eta_i =$$

$$= \begin{cases} \beta_0 + \beta_1 \log(age_i) & \text{for breath}_i = B \\ \beta_0 + \beta_2 + (\beta_1 + \beta_3)\log(age_i) & \text{for breath}_i = N \end{cases}$$

. Matching:

(intercept)	$\hat{\beta}_0$
breathN	$\hat{\beta}_2$
logage	$\hat{\beta}_1$
breathN:logage	$\hat{\beta}_3$

c) Model 2 fits. Res. Dev = 14.996 $\sim \chi^2_{14}$ if Model 2 fits. $\chi^2_{\alpha,14} \approx 23$ for $\alpha = 0.05$ from Cambridge Tables.

d)

Hypothesis $\Rightarrow H_0 : \beta_3 = 0$

Test statistic : $\frac{D_1 - D_2}{\phi} = D_1 - D_2 = 40.241 - 14.996$

$= 25.245$

Distribution under H_0

$$\frac{D_1 - D_2}{\phi} \overset{H_0}{\sim} \chi^2_{(n-p)-(n-q)}$$

p = # params in Model 1 = 3

q = # params in Model 2 = 4

$\phi = 1 \quad \therefore \quad D_1 - D_2 \overset{H_0}{\sim} \chi^2_1$.

Policy : reject H_0 iff $D_1 - D_2 > \chi^2_{\alpha,1}$

Equivalently we can use the p-value < 0.00000051 which is highly significant at the standard levels (0.01, 0.001, 0.05).

$\therefore$ There is very strong evidence that H_0 is false i.e. that $\beta_3 \neq 0$.

e) We might think to first reject Model 1 as fitting since Res. Dev = 40.2 for model 1, which is extremely large relative to the χ^2_{15} distribution. However rejecting Model 1 does not say anything specific about β_3. So a better idea is to use the z-score associated with $\hat{\beta}_3$: 4.957, which is very large relative to $N(0,1)$.

f) $$\log\left[\frac{\pi(z,B)}{1-\pi(z,B)} \div \frac{\pi(z,N)}{1-\pi(z,N)}\right] = \eta(z,B) - \eta(z,W) =$$

$$= \begin{cases} -\beta_2, & \text{Model 1} \\ -\beta_2 - \beta_3 \log(z), & \text{Model 2} \end{cases}.$$