$$VG \Rightarrow \beta_j$$

$$G + VG \Rightarrow \alpha_i + \beta_j$$

$$Age * (Gender + VG)$$

$$= Age + (Gender + VG) + Age.(Gender + VG)$$

$$= \beta_0 + \beta_1 x + (\alpha_i + \beta_j) + (\beta_0 + \beta_1 x)(\alpha_i + \beta_j)$$

$$= \cancel{\beta_0} + \cancel{\beta_1 x} + \alpha_i + \beta_j + \alpha_i \beta_0 + \beta_0 \beta_j + \alpha_i \beta_1 x + \beta_j \beta_1$$

$$= \alpha_i + \beta_j + (\alpha_i x + \beta_j x)$$

$$= \underline{\alpha_i + \beta_j + \alpha_i x + \beta_j x}$$

Eg.: Bin

link function $g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$

Linear predictor $\eta = \alpha_i + \beta_1 N + \beta_2 S$

$$g(\mu) = \eta$$

$$\therefore \ln\left(\frac{\mu}{1-\mu}\right) = \eta$$

$$\therefore \frac{\mu}{1-\mu} = e^{\eta}$$

$$\therefore \mu = e^{\eta} - \mu e^{\eta}$$

$$\therefore \mu = \frac{e^{\eta}}{1+e^{\eta}} \quad (1)$$

* <u>Model fitting</u>

x

* Model fitting

→ Use MLE to derive estimates of parameters

Eg: $X_i \sim Exp\left(\frac{1}{\mu_i}\right)$

$\therefore f(x_i) = \frac{1}{\mu_i} e^{-x_i/\mu_i}$

$\therefore f(x_i) = \exp\left[\ln\left(\frac{1}{\mu_i}\right) - \frac{x_i}{\mu_i}\right]$

$L = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n}$ ''

$\therefore \ln L = \sum \ln\left(\frac{1}{\mu_i}\right) - \sum \frac{x_i}{\mu_i}$

Link f^n for exp $= \frac{1}{\mu_i}$

Linear predictor $\eta = \alpha + \beta z_i$ [Given]

$\therefore g(\mu) = \eta$

$\therefore \frac{1}{\mu_i} = \alpha + \beta z_i$

$\therefore \ln L = \sum \ln(\alpha + \beta z_i) - \sum x_i(\alpha + \beta z_i)$

Diff w.r.t α –

$\therefore \frac{d \ln L}{d\alpha} = \sum \frac{1}{\hat{\alpha} + \hat{\beta} z_i} - \sum x_i = 0$

Diff w.r.t β –

$\therefore \frac{d \ln L}{d\beta} = \sum \frac{z_i}{\hat{\alpha} + \hat{\beta} z_i} - \sum x_i z_i = 0$

Linear

$\eta = \alpha + \beta$

Link f^n

$g(\mu) =$

$\therefore \frac{1}{\mu} = \alpha +$

pred

$\beta_1 x_1 + \beta_2 x_2^2$

$\beta_1 x_1 + \beta_2 x_2^2$

* Saturated Model

As no. of parameters ↑, error ↓

Saturated model is a model where fitted values = Observed values i.e. $\hat{\mu}_i = y_i$

$\Rightarrow$ No. of parameters = No. of observed values

Scaled Deviance

$SD = 2(l_s - l_m)$

l_s : log likelihood of saturated model

l_m : " " " current "

Eg: Under $Exp(\lambda)$ distⁿ —

$\ln L = -\sum \frac{y_i}{\mu_i} + \sum \ln\left(\frac{1}{\mu_i}\right)$

Under L_s, replace μ_i with y_i

Under l_m, " " " $\alpha + \beta x_i$

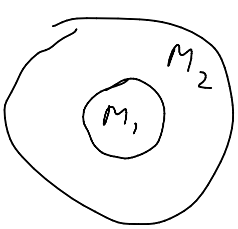
* Likelihood Ratio Test

H_0 : Model 1 is appropriate

H_1 : " 2 " " i.e addl parameter is significant

Condition: Models have to be nested & $M_1 < M_2$

SD No. of par



M_2
M_1

$M_1 \quad S_1 \quad q$

$M_2 \quad S_2 \quad p$

$S_1 > S_2 \quad p > q$

Test statistic : $\chi^2_{df,\alpha}$

$\chi^2_{cal} = S_1 - S_2$

Tabulated value : $\chi^2_{p-q,\,\alpha}$

At 5% LOS, $\chi^2_{tab} \doteq 2(p-q)$

D.C : Rej H_0 @ α% LOS if $\chi^2_{cal} > \chi^2_{tab}$

Roughly, $(S_1 - S_2) > 2(p-q)$

Q.

[Total 4]

12.4 An insurer wishes to use a generalised linear model to analyse the claim numbers on its motor portfolio. It has collected the following data on claim numbers y_i, $i = 1, 2, \ldots, 35$ from three different classes of policy:

Exam style

Class I	1	2	0	2	1	0	0	2	2	1
Class II	1	0	1	1	0					
Class III	0	0	0	0	0	1	0	1	0	0
	1	0	1	0	0	0	0	0	0	0

$5 \Rightarrow 2$

$10 \Rightarrow 1$

$21 \Rightarrow 0$

For these data values:

$$\sum_{i=1}^{10} y_i = 11 \qquad \sum_{i=11}^{15} y_i = 3 \qquad \sum_{i=16}^{35} y_i = 4$$

The company wishes to use a Poisson model to analyse these data.

(i) Show that the Poisson distribution is a member of the exponential family of distributions. [2]

(ii) The insurer decides to use a model (Model A) for which:

$$\log \mu_i = \begin{cases} \alpha & i = 1, 2, \ldots, 10 \\ \beta & i = 11, 12, \ldots, 15 \\ \gamma & i = 16, 17, \ldots, 35 \end{cases}$$

where μ_i is the mean of the relevant Poisson distribution. Derive the likelihood function for this model, and hence find the maximum likelihood estimates for α, β and γ. [4]

(iii) The insurer now analyses the simpler model $\log \mu_i = \alpha$, for all policies. Calculate the maximum likelihood estimate for α under this model (Model B). [2]

(iv) Show that the scaled deviance for Model A is 24.93, and calculate the scaled deviance for Model B. [5]

It can be assumed that $f(y) = y \log y$ is equal to zero when $y = 0$.

(v) Compare Model A directly with Model B, by calculating an appropriate test statistic. [2]

[Total 15]

ii) $\log \mu_i = \begin{cases} \alpha & ; \; i = 1, \dots, 10 \\ \beta & ; \; i = 11, \dots, 15 \\ \gamma & ; \; i = 16, \dots, 35 \end{cases}$

$\frac{e^{-\mu} \cdot \mu^{x}}{x!} = \exp$

$\ln \mu = \alpha \Rightarrow \mu = e$

$$L = \prod_{i=1}^{10} \exp\left[y_i(\alpha) - e^{\alpha} - \ln(y_i!)\right] \times \prod_{i=11}^{15} \exp\left[y_i(\beta) - e^{\beta} - \ln y_i!\right]$$

$$\times \prod_{i=16}^{35} \exp\left[y_i(\gamma) - e^{\gamma} - \ln(y_i!)\right] = \exp\left[\Sigma y_i(\alpha) - e^{10\alpha} -\right.$$

$\therefore \ln L = 11\alpha - 10e^{\alpha} + 3\beta - 5e^{\beta} + 4\gamma - 20e^{\gamma} - \Sigma \ln y!$

$\therefore \frac{\partial \ln L}{\partial \alpha} = 11 - 10e^{\hat{\alpha}} = 0$

$\therefore \hat{\alpha} = \ln\left(\frac{11}{10}\right) = \underline{0.09531}$

Similarly, $3 - 5e^{\hat{\beta}} = 0 \Rightarrow \hat{\beta} = \underline{-0.51083}$

$4 - 20e^{\hat{\gamma}} = 0 \Rightarrow \hat{\gamma} = \underline{-1.60944}$

iii) Model B : $\log \mu_i = \alpha$

$\therefore \ln L = 18\alpha - 35e^{\alpha} - \Sigma \ln y_i$

$\therefore \frac{d \ln L}{d\alpha} = 18 - 35e^{\hat{\alpha}} = 0$

$\therefore \hat{\alpha} = \underline{-0.66498}$

iv) $f(y) = y \log y = 0 \; ; \; y = 0$

$$\left[-\mu + x \ln \mu - \ln x!\right]$$

α

$$e^{x} = a$$

$$\ln e^{x} = \ln a$$

$$x = \ln a$$

$$\Sigma \ln y!\,]$$

$SD = 2(L_s - L_m)$

For saturated model —

$\ln L = \sum y_i \ln \mu_i - \sum \mu_i - \sum \ln y_i!$

Under a saturated model,

$\mu_i = y_i$

$\therefore \ln L = \underline{\sum y_i \ln y_i} - \sum y_i - \sum \ln y_i!$

$= 2\ln 2 \times 4 + 1\ln 1 \times 10 + 0\ln 0 \times 21$

$- 18 - 10 \times \ln 1! - 4 \times \ln 2! - 21 \ln 0!$

$:= -15.22741$

For Model A —

$\ln L = 1\alpha + 3\beta + 4\gamma - 10e^{\alpha} - 5e^{\beta} - 20e^{\gamma} - 4\ln 2$

Substitute parameter values from ii)

$\therefore \ln L = -27.69441$

$\therefore SD = 2(-15.22741 - (-27.69441)) = \underline{24.93}$

For Model B —

$\ln L = 18\alpha - 35e^{\alpha} = \underline{-32.74216}$

$\therefore SD = 2(-15.22741 + 32.74216) = \underline{35.0295}$

v.) $SD_B - SD_A = \underline{10.0995}$

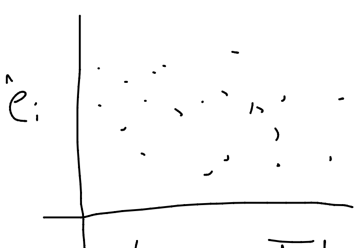
($\because$ Model B has fewer par compared to Model A)

$\chi^2_{p-q} = \chi^2_{3-1,\,5\%} = \chi^2_{2,\,5\%} = \underline{5.991}$

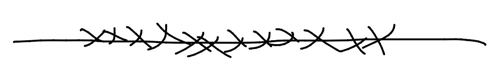
$\therefore$ $SD_B - SD_A > \chi^2_{2,\,5\%}$

$\therefore$ Rej H_0 @ 5% LOS & conclude Model A is a better fit

* Residual analysis



Uncorrelated



Normally distributed

Method 1 : Pearson residuals

$$\hat{e}_i = \frac{y_i - \hat{\mu}_i}{\sqrt{Var(\hat{\mu}_i)}}$$

standardised errors

Method 2 : Deviance residuals

$\hat{e}_i$ = sign of $(y_i - \hat{\mu}_i) \times d_i$

$SD = \sum d_i^2$

$SD = \sum d_i^2$

$d_i^2 \Rightarrow$ Contribution of i^{th} value to SD

$d_i = \sqrt{d_i^2}$

Eg: $\ln L_s = -\sum y_i + \sum y_i \ln y_i - \sum \ln y_i!$

$\therefore \ln L_{s_1} = -y_1 + y_1 \ln y_1 - \ln y_1!$

$\therefore \ln L_{M_1} = -e^{\alpha} + \alpha y_1 - \ln y_1!$

$\therefore d_1^2 = 2[\ln L_{s_1} - \ln L_{M_1}]$

$SD = 2(l_s - l_m)$

l_s = log-likelihood

1. Distⁿ & Linear predictor given
2. Identify link function basis the distⁿ given
3. Equate LP = LF
4. Arrive at likelihood fn
5. Replace link fn in likelihood eqn with LP (2)
6. Follow steps for MLE

A small insurer wishes to model its claim costs for motor insurance using a simple generalised linear model based on the three factors:

$$YO_i = \begin{cases} i=1 & \text{for 'young' drivers} \\ i=0 & \text{for 'old' drivers} \end{cases}$$

$$FS_j = \begin{cases} j=1 & \text{for 'fast' cars} \\ j=0 & \text{for 'slow' cars} \end{cases}$$

$$TC_k = \begin{cases} k=1 & \text{for 'town' areas} \\ k=0 & \text{for 'country' areas} \end{cases}$$

The insurer is considering three possible models for the linear predictor:

Model 1: $YO + FS + TC$

Model 2: $YO + FS + YO.FS + TC$

Model 3: $YO * FS * TC$

(i) Write each of these models in parameterised form, stating how many non-zero parameter values are present in each model. [6]

(ii) Explain why Model 1 might not be appropriate and why the insurer may wish to avoid using Model 3. [2]

i) Model 1: $YO + FS + TC$

$\alpha_i + \beta_j + \gamma_c$

$\alpha_Y, \alpha_O, \beta_F, \beta_S, \gamma_T, \gamma_C = 6$

Model 2: $YO + FS + TC + YO.FS$

$= \underbrace{YO + FS + YO.FS}_{YO*FS} + TC$

$= YO*FS + TC$ (2

$= \alpha_{ij} + \beta_k$

Model 3: $YO * FS * TC$

$= \underline{\underline{\alpha_{ijk}}}$

YFT YFC

YST YSC

$\alpha_{YFT}, \alpha_{YFC}, \ldots, \underline{\underline{8}}$

categorical $m-1$

cat * cat $m*n$

numerical 1

(least) Constant model $\underline{1}$

(Most) Saturated model No. of par = No. of data

Eg: Age (num) G (3,) 1 (U.R)

$= 1 + (2-1) + (2-1) + (2-1)$

$= \underline{4 \text{ par}}$

$* 2 + 1) = \underline{5 \text{ par}}$

Points

Eg: Age(num), G(3 v), L(U, R)

categorical

$y = \alpha$ ⇒ Null model or constant model =

$y = \alpha_i$ ⇒ α_m, α_F, α_o ⇒ No. of par = 3

$(m - 1)$

No. of cat in the variable

YO + FS + TC

O ⇒ 0

S ⇒ 0

C ⇒ 0

YSC ⇒ $y = \alpha_i$

OSC ⇒ $y =$

1