2. (3.5 points)

An actuary at a private passenger auto insurance company wishes to use a generalized linear model to create an auto frequency model using the data below.

Number of Claims

Gender	Territory A	Territory B
Male	700	600
Female	400	420

Number of Exposures

Gender	Territory A	Territory B
Male	1,400	1,000
Female	1,000	1,200

The model will include three parameters: β_1, β_2, and β_3, where β_1 is the average frequency for males, β_2 is the average frequency for Territory A, and β_3 is an intercept.

a. (0.5 point)

Define the design matrix [X].

b. (0.25 point)

Define the vector of responses [Y].

c. (2.25 points)

Assuming $\beta_3 = 0.35$, solve a generalized linear model with a normal error structure and identity link function for β_1.

d. (0.5 point)

The actuary determines that the analysis results would be improved by assuming a Poisson error structure with a log link function. Identify two reasons this structure may better suit this data.

CONTINUED ON NEXT PAGE

Question 2:

Model Solution 1

a) $[x] = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$

b) $[y] = \begin{pmatrix} 700/1400 \\ 600/1000 \\ 400/1000 \\ 420/1200 \end{pmatrix}$

$= \begin{bmatrix} 0.5 \\ 0.6 \\ 0.4 \\ 0.35 \end{bmatrix}$

c) Classical linear model

$0.5 = \beta_1 + \beta_2 + 0.35 + \varepsilon_1$	$\varepsilon_1 = 0.15 - \beta_1 - \beta_2$
$0.6 = \beta_1 + 0.35 + \varepsilon_2$	$\varepsilon_2 = 0.25 - \beta_1$
$0.4 = \beta_2 + 0.35 + \varepsilon_3$	$\varepsilon_3 = 0.05 - \beta_2$
$0.35 = .35 + \varepsilon_4$	$\varepsilon_4 = 0$

$SSE = (0.15 - \beta_1 - \beta_2)^2 + (0.25 - \beta_1)^2 + (0.05 - \beta_2)^2$

$dSSE/d\,\beta_1 = -2(0.15 - \beta_1 - \beta_2) - 2(0.25 - \beta_1) = 0$
$.8 = 4\beta_1 + 2\beta_2$
$\beta_1 = 0.2 - 0.5\beta_2$
$\beta_1 = 0.2 - 0.5(.1 - .5\beta_1)$
$\beta_1 = 0.2$

$dSSE/d\,\beta_2 = -2(0.15 - \beta_1 - \beta_2) - 2(0.05 - \beta_2) = 0$
$.4 = 2\beta_1 + 4\beta_2$
$B_2 = .1 - .5\beta_1$

d) Frequency is non-negative, so normal distribution is a poor fit.
A multiplicative relationship fits frequency better than additive relationship.

Model Solution 2

Average Frequencies = # claims/ # Exposures

Gender	Territory A (β_2)	Territory B
Male (β_1)	0.5	0.6
Female	0.4	0.35

Intercept (β_3)	Male (β_1)	Territory A (β_2)	Y
1	1	1	0.5
1	1	0	0.6
1	0	1	0.4
1	0	0	0.35

a) $[x] = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$

b) $y = \begin{bmatrix} 0.5 \\ 0.6 \\ 0.4 \\ 0.35 \end{bmatrix}$

$$\begin{bmatrix} 0.5 \\ 0.6 \\ 0.4 \\ 0.35 \end{bmatrix}$$

c) $\beta_3 = 0.35$

Normal error structure, identity link -> $g(x) = x$; $g^{-1}(x) = x$

$$\mu = E[Y] = \begin{pmatrix} \beta_1 + \beta_2 + \beta_3 \\ \beta_1 + \beta_3 \\ \beta_2 + \beta_3 \\ \beta_3 \end{pmatrix}$$

Identify likelihood function:

$$L(y;\mu,\sigma^2) = \prod_{i=1}^{n} \exp\{-\frac{(y_i - \mu_i)^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\}$$

Take the logarithm to convert the product of many terms into a sum:

$$l(y;\mu,\sigma^2) = \sum_{i=1}^{n} -\frac{(y_i - \sum_{j=1}^{p} X_{ij}.\beta_j)^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)$$

$\mu = \Sigma x_{ij}\, \beta_j$

*Ignoring constant term $1/2\ln(2\pi\sigma^2)$

$$\to l(y, \mu, \sigma^2) = -\frac{(0.5 - (\beta_1 + \beta_2 + \beta_3))^2}{2\sigma^2} - \frac{(0.6 - (\beta_1 + \beta_3))^2}{2\sigma^2} - \frac{(0.4 - (\beta_2 + \beta_3))^2}{2\sigma^2} - \frac{(0.35 - (\beta_3))^2}{2\sigma^2}$$

Maximize the logarithm of the likelihood function by taking partial derivatives with respect to each covariate setting equation to 0 and solving system of equations. Question only asks to solve for β_1.

$$\frac{dl(y, \mu, \sigma^2)}{d\beta_1} = -\frac{2(-1)(0.5 - \beta_1 - \beta_2 - \beta_3)}{2\sigma^2} - \frac{2(-1)(0.6 - \beta_1 - \beta_3)}{2\sigma^2} = 0$$

➔ $0.5 - \beta_1 - \beta_2 - \beta_3 + 0.6 - \beta_1 - \beta_3 = 0 \to 0.4 = 2\beta_1 + \beta_2$

$$\frac{dl(y, \mu, \sigma^2)}{d\beta_2} = -\frac{2(-1)(0.5 - \beta_1 - \beta_2 - \beta_3)}{2\sigma^2}) - \frac{2(-1)(0.4 - \beta_2 - \beta_3)}{2\sigma^2} = 0$$

➔ $0.5 - \beta_1 - \beta_2 - \beta_3 + 0.4 - \beta_2 - \beta_3 = 0 \to 0.2 = \beta_1 + 2\beta_2$

$0.4 = 2\beta_1 + \beta_2$
$0.2 = \beta_1 + 2\beta_2$

$\to 0.8 = 4\beta_1 + 2\beta_2 \to 2\beta_2 = 0.8 - 4\beta_1$
$\to 0.2 = \beta_1 + 0.8 - 4\beta_1$
$\to 0.6 = 3\beta_1$
$\to 0.2 = \beta_1$

d) The Poisson structure is more appropriate for this data use for a frequency model because:

a. Values for freq. are restricted to positive values; normality violates this assumptions
b. Normality assumes a fixed variance while Poisson structure allows the variance to increase with mean (more weight to observations at left of distribution)

Examiner's Comments
****
Part a
For full credit, the candidate must correctly populate a design matrix **X** based on the information given in the problem. There were multiple correct solutions to this problem, dependent on the order of the βs the candidate chose.

The most common error was omitting or incorrectly representing the intercept term in the matrix.

Part b
For full credit, the candidate must correctly populate the vector of responses **Y** based on the information given in the problem. The answer to part b is dependent on the order of βs the candidate chose in part a.

The most common error was providing the exposures instead of the frequencies as the values for **Y.**

Part c
For full credit, the candidate must solve the GLM by recognizing that the normal error structure and identity link function is identical to a classical linear model. The correct approach involved setting up linear equations, minimizing the sum of squared errors by taking the partial derivatives of each β, and solving the system of equations to arrive at $\beta_1 = 0.2$. Full credit was also given to candidates who did not identify that a normal error structure and identity link function was equivalent to a CLM, but solved by using the appropriate likelihood function, recognizing that solving the log-likelihood was equivalent.

A common error was to differentiate for β_3, which was a given in the problem, and attempting to solve the system of 3 equations.

Part d
For full credit, the candidate must identify two unique reasons that the Poisson error structure and/or a log link error function are a better fit for the data.
The most common incorrect or no credit response was "Poisson is a common distribution for frequencies", as this response did not show an understanding of *why* this was the case.
****