2. (2.25 points)

A private passenger auto insurance company orders a report whenever it writes a policy, showing what other insurance the policyholder has purchased. The following table shows claim frequencies (per 100 earned car-years) for bodily injury liability coverage, split by whether the policyholder has a homeowners policy and whether the policyholder had a prior auto policy:

	Homeowners Policy	
Prior Auto Policy	Yes	No
Yes	3	5
No	8	12

The table does not include the experience of policyholders with missing data.

a. (1.25 points)

Specify the following structural components of a generalized linear model that estimates frequencies for this book of business.

i. Error distribution
ii. Link function
iii. Vector of responses
iv. Vector of model parameters
v. Design matrix

b. (1 point)

Describe how the missing data may cause problems for the company in developing the model, and suggest a solution.

CONTINUED ON NEXT PAGE

Question 2:

Model Solution 1

a) i. Error should be Poisson for frequency.

$$P(x = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

ii. Link function should be log link for multiplicative model.

$$g(x) = \ln(x)$$
$$g^{-1}(x) = e^x$$

iii. $\begin{bmatrix} 3 \\ 5 \\ 8 \\ 12 \end{bmatrix}$

iv. $\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$ where β_0 is intercept

v.

$$\begin{matrix} & x_1 & x_2 \end{matrix}$$
$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

x_1 is Prior Auto Policy = Yes

x_2 is Homeowners Policy = Yes

b) Missing data can be problematic. If you put "unknown" as a level for each factor, for example, they will be perfectly correlated with each other. This will cause aliasing. To solve, you can eliminate the level from one of the factors so there are no linear dependencies. If there are linear dependencies, there will be no unique solution in beta parameters and any arbitrary amount can be added to one parameter and subtracted from the other.

Model Solution 2

a) i) Error distribution → Poisson (since modeling claim frequencies)

ii) Link fn → Log link $E(\underline{\gamma}) = \underline{\mu}$ $g^{-1}(\underline{\eta}) = e^{\underline{\eta}}$

iii) $\underline{Y}=(3, 5, 8, 12)^T$

iv) $\underline{\beta} =(\beta_1, \beta_2, \beta_3)^T$

→ let $x_1 = \begin{cases} 1 & \text{prior auto policy = Yes} \\ 0 & \text{otherwise} \end{cases}$

$x_2 = \begin{cases} 1 & \text{prior auto policy = No} \\ 0 & \text{otherwise} \end{cases}$

$x_3 = \begin{cases} 1 & \text{Homeowners policy = Yes} \\ 0 & \text{otherwise} \end{cases}$

v) $X = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}$ → ignoring policyholders with missing data

b) Missing data can lead to extrinsic aliasing. This occurs when there are linear dependencies in the observed data because of the nature of the data. In this case, the "missing" level for "prior auto policy" will be perfectly correlated with the "missing" level for "homeowners policy". This can lead to convergence problems or confusing results. A solution would be to exclude these missing data records, or to reclassify them to an appropriate level.

Examiner's Comments

****

Part a

The most common errors were improperly identifying the error and link functions, or proposing a suboptimal alternative answer without proper justification. Answers containing design matrices that didn't correctly correspond to the response vector or did not correct for aliasing were also encountered.

Generally, candidates did well with identifying the error distribution. Candidates who got the error distribution wrong either left it blank or picked a non-Poisson distribution without justifying that choice. Very few appeared to confuse the error distribution with the link function. The less-prepared candidate could usually guess at a distribution for the errors and would often name a distribution for the link function as well. Candidates who specified an incorrect function usually gave the identity function without justification.

Most candidates got at least two thirds of the vectors and design matrix correct. When they lost credit it was typically because they didn't label the vectors clearly

enough. For example, a candidate might list all three without assigning subparts or labeling, or list a model vector without associating the betas with anything. Another common mistake was to specify a design matrix inconsistent with the vectors, usually because the matrix values were flipped between the y=5 and y=8 cases.

Part b

Most candidates got partial credit on this. By far the most typical mistake was an omission: Usually a candidate would either identify an explanation of how aliasing was a problem with the data, or how it would impact the model, but not both. Most candidates were able to present a reasonable solution or workaround for the problem. Some candidates lost credit by contradicting a correct statement, typically by implying that aliasing was a desirable characteristic of a model.