4. (3 points)

An actuary is helping design a new internet liability product that would use industry as a rating factor. Different business types such as restaurants, auto manufacturers, and dairy farms would fall into different industry groups. The actuary wants to create several industry factors from a combination of insurance and demographic data, and use this to classify business types into industry groups.

a. (1 point)

Describe two reasons that a generalized linear model might not be appropriate for developing industry factors.

b. (0.25 point)

Describe a benefit that a principal component method would have over a generalized linear model for determining the industry factors.

c. (1 point)

Briefly describe the major steps in using a cluster analysis to group the industry factors.

d. (0.75 points)

Describe two test statistics that could be used to determine the optimal number of groups from the cluster analysis. Identify which statistic would be preferred when variables are correlated.

CONTINUED ON NEXT PAGE

**Question 4:**

***Part a***

***Model Solution 1***

- There would likely be Intrinsic Aliasing, since you are using qualitative variables.
- It is not apparent what type of error structure or link function should be used for Industry type.

***Model Solution 2***

- GLM assumes that each obs is independent of each other which is probably not the case here as restaurants are clearly affected by dairy farms.
- The observations may not come from exponential family distribution.

**Examiner's Comments**

**********

Candidates generally performed poorly on this subpart. A common response was that GLMs could not be used for grouping business types into industries. However, this question was not focused on that grouping task but rather on the deficiencies in the GLM assumptions as they relate to using a complex data set (industry and demographic data) when modeling response variables that differ from actuarial norms (frequency, severity, claims).

Examples of other acceptable responses include:

- GLMs are vulnerable to aliasing. If aliasing occurs, convergence of model results may be difficult. Aliasing can be intrinsic (category design), extrinsic (nature of data) or near. In this case, actuary is combining demographic data that is likely to be incomplete across risks and may lead to some form of aliasing.
- Need to select error and link functions. In general, actuaries have an idea of error functions for certain things (severities, claims) but would not have a good place to start with this data. Also, there might not be any functions that this kind of data well.
- Hard to identify the error structure and link function---we don't have a feel if the exponential family is a good model for these factors.
- There could be issues in the data (missing an industry code, etc.) that could lead to aliasing or near aliasing, where all the missing data values are correlated.

**********

***Part b***

***Model Solution1***

The principal component (analysis) will identify the representative variable and determine the most significant factors---maximizing the proportion of total variance explained.

***Model Solution 2***

Principal component analysis identifies variables that are most predictive of [the] outcome, allowing one to eliminate other correlated variables from the model making the model simpler without much loss of function.

**Examiner's Comments**

**********

Candidates generally performed poorly on this subpart. We were looking for answers relating to reducing dimensionality of the explanatory variables or accounting for correlation between variables/finding the most predictive variables.

**********

***Part c***

***Model Solution 1***

1) Select the number of groups for k-means clustering using the methods in part (d)
2) Randomly assign factors to groups
3) Compute average factor for each group, called the centroid
4) Calculate distance of each factor to centroids, re-assigning to the closest centroid
5) If any factors changed groups, go back and repeat at step 3 with new groups until the algorithm stabilizes

***Model Solution 2***

1) Choose the number of clusters
2) Randomly assign risks to clusters
3) Compute centroid of each cluster and assign each risk to the closest centroid based on euclidean mean criteria.
4) If any of the risks move to a different cluster, repeat step 3.

**Examiner's Comments**

**********

Candidates generally performed well on this subpart.

**********

***Part d***

***Model Solution 1***

Calinski-Harbasz statistic: Measures the between variance of the clusters divided by the within variance.

Cubic Clustering Criterion (CCC): Measures variance explained by the clusters compared to clusters formed at random according to a multi-dimension uniform distribution.

If correlation is present, CCC performs worse, so use Calinski-Harbasz Statistic.

***Model Solution 2***

Calinski and Harbasz test statistic

$\frac{\text{Trace (B)}}{(k-1)}$
$\frac{\text{Trace (W)}}{(n-k)}$

Preferred when variables are correlated

Cubic Clustering Criterion
Compares variance explained by clusters to that explained by randomly assigned clusters.

**Examiner's Comments**

**********

Candidates generally performed well on this subpart. The most common mistakes when describing the CCC statistic were failing to mention variance or comparing cluster variance with "total" variance instead of variance of a "random group". Some candidates either failed to mention which statistic was preferred or indicated that CCC was "less" preferred.

**********