MA409: Statistical Data Analysis (SAS)

Assignment 5 (May 25 – June 15)

Note: Please work on 5.4-5.6 with SAS procedures.

5.1 The PMF of the negative binomial distribution $NB(r, p)$ is:

$$f(y; r, p) = \Pr(Y = y) = \binom{y + r - 1}{y}(1 - p)^r p^y.$$

Show that

(1) $NB(r, p)$ belongs to the exponential family. (5 points)

(2) For $Y \sim NB(r, p)$, compute $E(Y)$ and $Var(Y)$. (10 points)

5.2 $Y_1, \dots, Y_n$ are independent and $Y_i \sim \text{Poisson}(\mu_i)$. Let M be a model of interest and $\hat{y}_i$ is the estimated value of Y_i under M (y_i is the observed value of Y_i).

(1) Show that the deviance of model M is $D = 2[\sum y_i \log(y_i/\hat{y}_i) - \sum(y_i - \hat{y}_i)]$. (5 points)

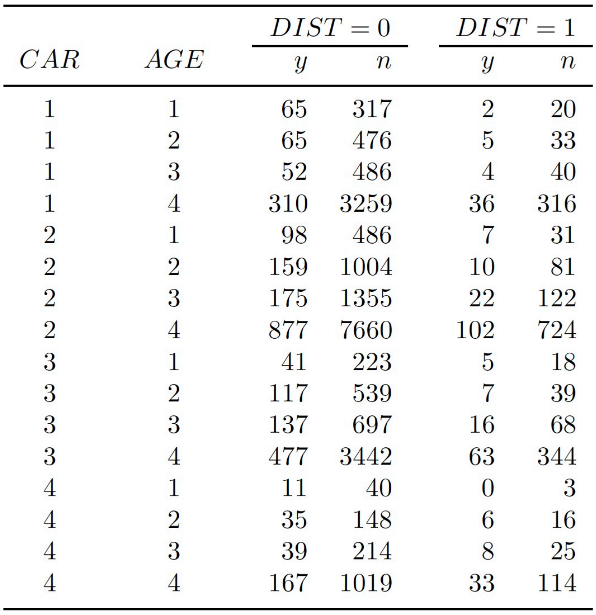
(2) Let $\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$. Show that the score statistic for β_0 is $U_0 = \sum(y_i - \mu_i)$. (5 points)

(3) Show that the deviance of model M simplifies to $D = 2\sum y_i \log(y_i/\hat{y}_i)$. (5 points)

5.3 Show that the Poisson distribution $\text{Poisson}(\mu)$ is the limit of a Binomial distribution $\text{Binomial}(n, p)$, for which the $p = \mu/n$ as n goes to infinity. (10 points)

5.4 For the HR dataset used in Problem 4.4 (2) in Assignment 4, we fit a logistic regression model on the probability that an employee would leave the company and the resulting AUC of the model is 0.8194. Think of a way to improve the model and obtaining a logistic regression model with AUC > 0.93. Hint: think of properly discretizing some of the continuous explanatory variables. (20 points)

5.5 The data in the following table are number of insurance policies (n) and number of claims (y) for cars in various insurance categories (CAR), tabulated by age group of policy holder (AGE), and district where the policy holder lived ($DIST = 1$ for London and other major cities, and $DIST = 0$ otherwise).



CAR	AGE	$DIST = 0$		$DIST = 1$	
		y	n	y	n
1	1	65	317	2	20
1	2	65	476	5	33
1	3	52	486	4	40
1	4	310	3259	36	316
2	1	98	486	7	31
2	2	159	1004	10	81
2	3	175	1355	22	122
2	4	877	7660	102	724
3	1	41	223	5	18
3	2	117	539	7	39
3	3	137	697	16	68
3	4	477	3442	63	344
4	1	11	40	0	3
4	2	35	148	6	16
4	3	39	214	8	25
4	4	167	1019	33	114

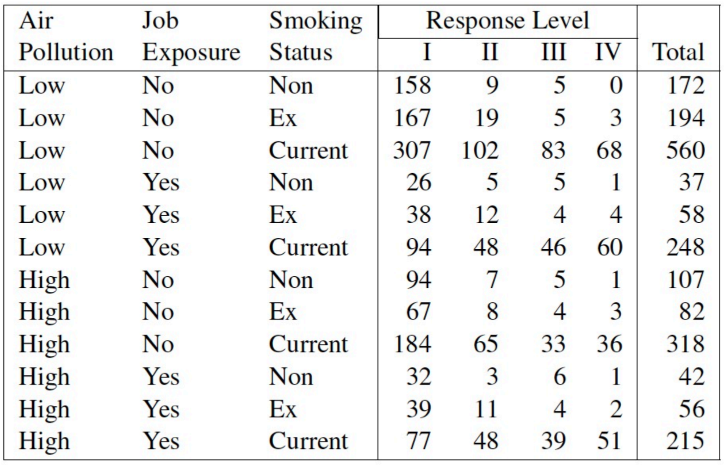
(1) Load the data into SAS so that it can be used to fit a Poisson regression model. (5 points)

(2) Fit a Poisson regression model on the rate of claim with CAR, AGE, $DIST$ as categorical explanatory variables, display and interpret your results (use 1 for CAR and AGE as the reference category, use 0 for $DIST$ as the reference category). (10 points)

(3) Fit a negative binomial regression model on the rate of claim with the same explanatory variables, compare your results with (2) and explain your findings. (5 points)

5.6 The data in the following table are from an epidemiological study of chronic respiratory disease. Researchers collected information on subjects' exposure to general air pollution, exposure to pollution in their jobs, and whether they smoke. The response measured was chronic respiratory disease status which had four categories:

- Level I: no symptoms
- Level II: cough or phlegm less than three months a year
- Level III: cough or phlegm more than three months a year
- Level IV: cough and phlegm plus shortness of breath more than three months a year



Air Pollution	Job Exposure	Smoking Status	Response Level				Total
			I	II	III	IV	
Low	No	Non	158	9	5	0	172
Low	No	Ex	167	19	5	3	194
Low	No	Current	307	102	83	68	560
Low	Yes	Non	26	5	5	1	37
Low	Yes	Ex	38	12	4	4	58
Low	Yes	Current	94	48	46	60	248
High	No	Non	94	7	5	1	107
High	No	Ex	67	8	4	3	82
High	No	Current	184	65	33	36	318
High	Yes	Non	32	3	6	1	42
High	Yes	Ex	39	11	4	2	56
High	Yes	Current	77	48	39	51	215

(1) What’s the name of the model you would choose to model the response? Provide your rationale of choosing that model. (5 points)

(2) Load the data into SAS so that it can be used to fit the model in (1). (5 points)

(3) Fit the model in (1) with the main effects of the three explanatory variables provided in the table. Provide evidence on whether the model is appropriate. Display and interpret your results. (10 points)