MA409: Statistical Data Analysis (SAS)

Assignment 5

Note: Please work on problem 3-5 using SAS procedures. Please provide the SAS code in a separate ***.sas file*** and the outputs from SAS (using screenshots) together with problems 1-2 in a single ***PDF file***. The SAS datasets are under *~/my_shared_file_links/u44964992/Assignments/* on SAS OnDemand for Academics.

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 (assume that $r > 0$ is known and $p \in [0, 1]$ is the parameter of interest)

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

(2) For $Y \sim NB(r, p)$, compute $E(Y)$, $\text{Var}(Y)$ and show that $\text{Var}(Y) = E(Y) + [E(Y)]^2/r$. (5 points)

(3) Derive the canonical link function of a generalized linear model assuming the negative binomial distribution for the response variable, i.e., a negative binomial regression model. (5 points)

(4) Provide a reason explaining why the log-link function is used more often for a negative binomial regression model than the canonical link function derived in (3)? (5 points)

2. Consider a random sample $Y_1, Y_2, \dots, Y_n$ from the exponential distribution (y_i is the observed value of Y_i):

$$f(y_i; \theta_i) = \theta_i \exp(-y_i \theta_i).$$

Derive the deviance by comparing the saturated model with different values of θ_i for each Y_i and the model with $\theta_i = \theta$ for all i. Justify that the deviance is always nonnegative. (10 points)

3. The table below shows the result of a sample of people on whether they report drinking alcohol frequently (Yes/No) and on the four scales of the Myers-Briggs personality test: Extroversion/Introversion (E/I, 外向/内向), Sensing/iNtuitive (S/N, 实感/直觉), Thinking/Feeling (T/F, 思考/情感), Judging/Perceiving (J/P, 判断/感知). The personality test results in 16 personality types: ESTJ, ESTP, ESFJ, ESFP, ENTJ, ENTP, ENFJ, ENFP, ISTJ, ISTP, ISFJ, ISFP, INTJ, INTP, INFJ, INFP.

Extroversion/Introversion		E				I			
Sensing/iNtuitive		S		N		S		N	
				Alcohol Frequently					
Thinking/Feeling	Judging/Perceiving	Yes	No	Yes	No	Yes	No	Yes	No
T	J	10	67	3	20	17	123	1	12
	P	8	34	2	16	3	49	5	30
F	J	5	101	4	27	6	132	1	30
	P	7	72	15	65	4	102	6	73

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

(2) Fit a logistic regression model on the probability of drinking alcohol frequently using the four scales as predictors (**Note:** use I, N, F, P as the corresponding reference level). Obtain the odds ratio estimates and interpret the odds ratios. Plot the predicted probabilities for the 16 personality types and state which personality type has the highest predicted probability of drinking alcohol frequently. (10 points)

(3) For the logistic regression model in (2), drop the Judging/Perceiving scale, and add the interaction effect between the Sensing/iNtuitive and Thinking/Feeling scales. Obtain the odds ratio estimates of Sensing vs. iNtuitive people for Thinking and Feeling people separately and interpret your results. (5 points)

4. A big company wants to understand why its employees are leaving the company. The data "*HRdata.csv*" for 14,999 employees are provided by the HR department including satisfaction level, latest evaluation (yearly), number of projects worked on, average monthly hours, time spent in the company (in years), work accident within the past 2 years (0-no, 1-yes), promotion within the last 5 years (0-no, 1-yes), and salary level (low, medium, high). The response variable of interest is whether the employee left the company (0-no, 1-yes).

(1) Use graphical methods to explore the possible differences in the explanatory variables between employees who left and who didn't leave the company. Interpret your findings. (10 points)

(2) Fit the following logistic regression model on the probability that an employee would leave the company:

$$\begin{aligned}\text{logit}(p) = \beta_0 &+ \beta_1 \text{satisfaction\_level} + \beta_2 \text{last\_evaluation} + \beta_3 \text{number\_project} \\ &+ \beta_4 \text{average\_monthly\_hours} + \beta_5 \text{time\_spent\_company} \\ &+ \beta_6 \text{I}(\text{work\_accident} = 1) + \beta_7 \text{I}(\text{promotion\_last\_5years} = 1) \\ &+ \beta_8 \text{I}(\text{salary} = \text{medium}) + \beta_9 \text{I}(\text{salary} = \text{high}).\end{aligned}$$

Obtain the odds ratio estimates and interpret the odds ratios. Plot the ROC curve of the model, obtain the AUC, and interpret the AUC. (10 points)

(3) Think of a way to improve the model in (2) and obtain a logistic regression model with AUC > 0.93. Hint: think of properly discretizing some of the continuous explanatory variables to categorical variables. (10 points)

5. The data in the following table are the number of children (y) ever born to married women of the Indian race (n) classified by duration since their first marriage (0-4 years, 5-9 years, 10-14 years, 15-19 years, 20-24 years, 25-29 years), educational level (none, lower primary, upper primary, and secondary or higher), and type of place of residence (Suva, other urban, and rural).

Duration	Education	Suva		Urban		Rural	
		y	n	y	n	y	n
0-4	None	4	8	14	12	60	62
0-4	LP	24	21	23	27	98	102
0-4	UP	38	42	41	39	104	107
0-4	SH	37	51	35	51	35	47
5-9	None	31	10	59	13	171	70
5-9	LP	80	30	98	37	317	117
5-9	UP	49	24	118	44	200	81
5-9	SH	38	22	48	21	47	21
10-14	None	49	12	75	18	364	88
10-14	LP	99	27	143	43	546	132
10-14	UP	58	20	105	29	197	50
10-14	SH	24	12	50	15	30	9
15-19	None	59	14	108	23	577	114
15-19	LP	153	31	225	42	481	86
15-19	UP	41	13	92	20	135	30
15-19	SH	11	4	19	5	2	1
20-24	None	118	21	118	22	756	117
20-24	LP	91	18	147	25	431	68
20-24	UP	47	12	65	13	132	23
20-24	SH	13	5	16	3	5	2

(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 birth with *Duration*, *Education*, *Residence* as categorical explanatory variables, display and clearly interpret your results (**Note:** use "0-4" for *Duration*, "None" for *Education*, "Suva" for *Residence* as the reference level, only consider the main effects, no need to include any interaction effect). (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)