MA409: Statistical Data Analysis (SAS)

Assignment 5 (May 23 – June 12)

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 s

(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)

Solution:

(1) The PMF of $NB(r,p)$ is:

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

Let $h(y) = \binom{y+r-1}{y}$, $\theta = \log(p) \Rightarrow p = e^\theta, A(\theta) = -r\log(1-p) = -r\log(1-e^\theta)$, and $\phi = 1$, then $f(y;r,p) = h(y)\exp\left\{\frac{y\theta - A(\theta)}{\phi}\right\}$. Therefore, by the definition of exponential family, $NB(r,p)$ belongs to the exponential family.

(2) For Y following an exponential family distribution, we have:

$$E(Y) = A'(\theta) \text{ and } Var(Y) = \phi A''(\theta).$$

Hence:

$$E(Y) = A'(\theta) = \frac{d(-r\log(1-e^\theta))}{d\theta} = \frac{re^\theta}{1-e^\theta} = \frac{rp}{1-p}.$$

$$Var(Y) = A''(\theta) = \frac{d^2(-r\log(1-e^\theta))}{d\theta^2} = \frac{re^\theta}{(1-e^\theta)^2} = \frac{rp}{(1-p)^2}.$$

5.2 $Y_1, \ldots, Y_n$ 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)

Solution:

(1) $Y_1, \ldots, Y_n$ are independent and $Y_i \sim \text{Poisson}(\mu_i)$, so the log-likelihood function is:

$$\ell(\boldsymbol{\beta}; \boldsymbol{y}) = \sum y_i \log \mu_i - \sum \mu_i - \sum \log(y_i!).$$

For the saturated model, μ_i's are all different, so $\boldsymbol{\beta} = (\mu_1, \ldots, \mu_n)^\top$. It's not difficult to obtain that the maximum likelihood estimates are $\hat{\mu}_i = y_i$, so the maximum value of the log-likelihood under the saturated model is

$$\ell(\hat{\boldsymbol{\mu}}_S) = \sum y_i \log y_i - \sum y_i - \sum \log(y_i!).$$

For a model of interest M, the maximum value of the log-likelihood is ($\hat{y}_i = \hat{\mu}_i$ because $E(Y_i) = \mu_i$)

$$\ell(\hat{\boldsymbol{\beta}}_M) = \sum y_i \log(\hat{y}_i) - \sum \hat{y}_i - \sum \log(y_i!).$$

Therefore, the deviance is

$$D = 2\left(\ell(\hat{\boldsymbol{\mu}}_S) - \ell(\hat{\boldsymbol{\beta}}_M)\right) = 2\left[\sum y_i \log\left(\frac{y_i}{\hat{y}_i}\right) - \sum (y_i - \hat{y}_i)\right].$$

(2) As $\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$, then $\mu_i = \exp\{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\}$, so:

$$\frac{\partial \log(\mu_i)}{\partial \beta_0} = 1 \; and \; \frac{\partial \mu_i}{\partial \beta_0} = \exp\{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\} = \mu_i.$$

Hence, the score statistic for β_0 is

$$U_0 = \frac{\partial \ell(\boldsymbol{\beta}; \boldsymbol{y})}{\partial \beta_0} = \sum y_i \frac{\partial \log(\mu_i)}{\partial \beta_0} - \sum \frac{\partial \mu_i}{\partial \beta_0} = \sum (y_i - \mu_i).$$

(3) From (2) we have that the score statistic for β_0 is $U_0 = \sum(y_i - \mu_i)$. The solution of β_0 is obtained by setting $U_0 = 0$, so that under model M,

$$\sum \hat{\mu}_i = \sum y_i.$$

Therefore, the deviance of model M is ($\hat{y}_i = \hat{\mu}_i$)

$$D = 2\left[\sum y_i \log\left(\frac{y_i}{\hat{y}_i}\right) - \sum (y_i - \hat{y}_i)\right] = 2\sum y_i \log\left(\frac{y_i}{\hat{y}_i}\right).$$

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)

Solution:

The PMF of $\text{Binomial}(n, p)$ is (let $p = \mu/n$)

$$f(y; p) = \Pr(Y = y) = \binom{n}{y} p^y (1-p)^{n-y} = \frac{n!}{y!\,(n-y)!} \left(\frac{\mu}{n}\right)^y \left(1 - \frac{\mu}{n}\right)^{n-y}$$

$$= \left[\frac{n!}{(n-y)!} \frac{1}{n^y} \left(1 - \frac{\mu}{n}\right)^{-y}\right] \times \left[\frac{1}{y!} \mu^y \left(1 - \frac{\mu}{n}\right)^n\right] = \text{Part I} \times \text{Part II}.$$

$$\text{Part I} = \frac{(n-y+1)\cdots n}{(n-\mu)^y} \xrightarrow{n\to\infty} 1.$$

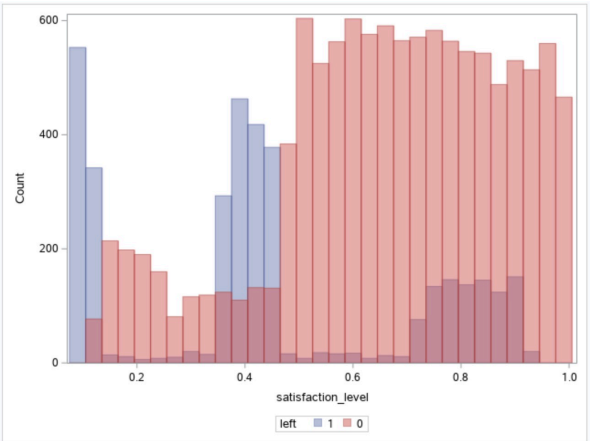
$$\text{Part II} = \frac{1}{y!} \mu^y \left(1 - \frac{\mu}{n}\right)^n \xrightarrow{n\to\infty} \frac{1}{y!} \mu^y e^{-\mu}.$$

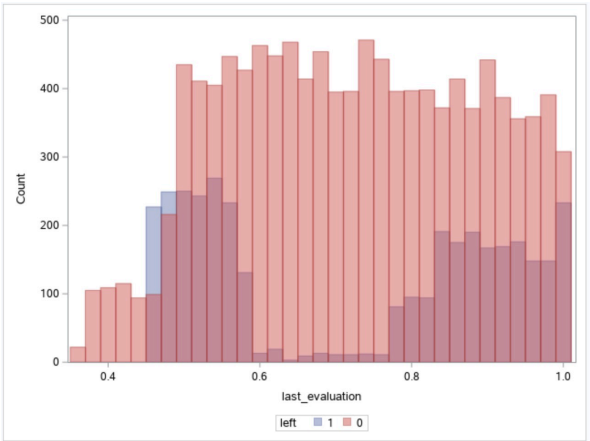
Therefore, $f(y; p) \xrightarrow{n\to\infty} \frac{1}{y!} \mu^y e^{-\mu}$ which is the PMF of $\text{Poisson}(\mu)$, i.e., $\text{Poisson}(\mu)$ is the limit of $\text{Binomail}(n, p)$ with $p = \mu/n$ as $n \to \infty$.

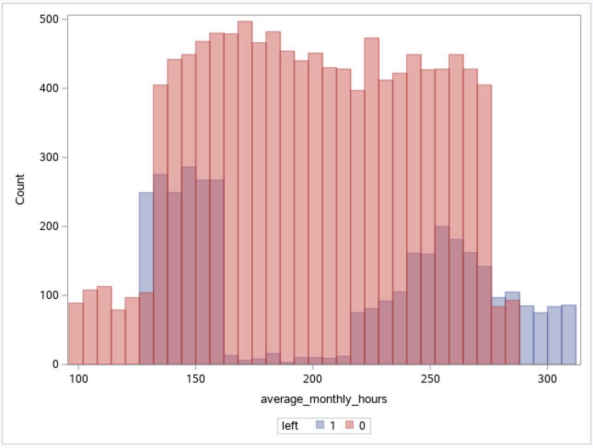
5.4 For the HR dataset used in Problem 4.4 (2) in Assignment 4, we fit a logistic regression model on the probability that an employee would leave the company and the resulting AUC of the model is 0.8194. Think of a way to improve the model and obtaining a logistic regression model with AUC > 0.93. Hint: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

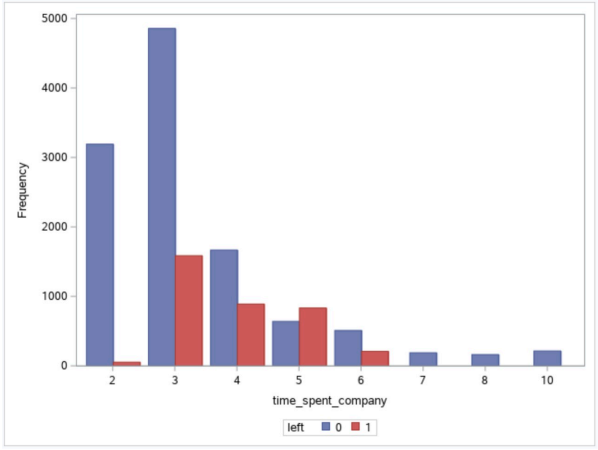
Solution: Solution is not unique. Below, I will provide my solution.

Look at the histograms of *satisfaction_level*, *last_evaluation*, *average_monthly_hours*, as well as the bar plot of *times_spent_company*:





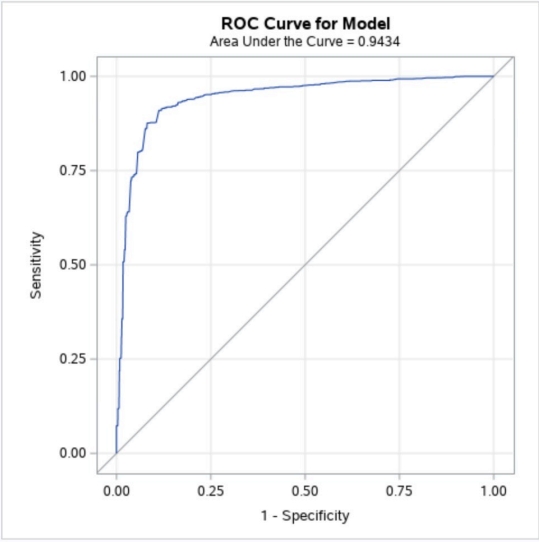




It is natural to discretize the four continuous variables as:

satisfaction_level		last_evaluation	
< 0.4	low	< 0.6	low
0.4 - 0.7	medium	0.6 - 0.78	medium
> 0.7	High	> 0.78	high
average_monthly_hours		**time_spent_company**	
< 165	low	2	short
165 - 220	normal	3 - 6	medium
220-275	high	>= 7	long
> 275	toohigh	---	---

Also treating *number_project* as a class variable, we refit the logistic regression model by replacing the four continuous variables with their corresponding discretized version, the resulting AUC is 0.9434 with the ROC Curve given below:



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).

		$DIST = 0$		$DIST = 1$	
CAR	AGE	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)

Solution: See "Assignment5.sas".

(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)

Solution: SAS program to fit the Poisson regression model is provided in "Assignment5.sas". The outputs are provided below:

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	24	23.7090	0.9879
Scaled Deviance	24	23.7090	0.9879
Pearson Chi-Square	24	22.3393	0.9308
Scaled Pearson X2	24	22.3393	0.9308
Log Likelihood		14129.7072	
Full Log Likelihood		-96.0346	
AIC (smaller is better)		208.0693	
AICC (smaller is better)		214.3302	
BIC (smaller is better)		219.7952	

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.8102	0.0753	-1.9578	-1.6626	577.61	<.0001
car	2	1	0.1623	0.0505	0.0633	0.2613	10.32	0.0013
car	3	1	0.3935	0.0550	0.2858	0.5013	51.22	<.0001
car	4	1	0.5654	0.0723	0.4237	0.7071	61.19	<.0001
age	2	1	-0.1890	0.0828	-0.3513	-0.0267	5.21	0.0225
age	3	1	-0.3421	0.0813	-0.5015	-0.1828	17.71	<.0001
age	4	1	-0.5327	0.0698	-0.6695	-0.3960	58.28	<.0001
dist	1	1	0.2185	0.0585	0.1038	0.3332	13.93	0.0002
Scale		0	1.0000	0.0000	1.0000	1.0000		

The deviance and Pearson Chi-square statistics divided by the degrees of freedom are close to 1, indicating no overdispersion and the Poisson regression model fits the data well.

The parameter estimates indicate that:

- Compared with cars in insurance category 1, cars in insurance category 2, 3, 4 have higher rate of claim and the rate tends to increase as the insurance category number increases.
- Compared with policy holders in age group 1, policy holders in age group 2, 3, 4 have higher rate of claim and the rate tends to increase as the age group number increases.
- Policy holders lived in London or other major cities tend to have higher rate of claim than those lived elsewhere.

(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)

Solution: SAS program to fit the negative binomial regression model is provided in "Assignment5.sas". The outputs are provided below:

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.8102	0.0753	-1.9578	-1.6626	577.59	<.0001
car	2	1	0.1623	0.0505	0.0633	0.2613	10.32	0.0013
car	3	1	0.3935	0.0550	0.2857	0.5013	51.22	<.0001
car	4	1	0.5654	0.0723	0.4237	0.7071	61.19	<.0001
age	2	1	-0.1890	0.0828	-0.3513	-0.0267	5.21	0.0225
age	3	1	-0.3421	0.0813	-0.5015	-0.1828	17.71	<.0001
age	4	1	-0.5327	0.0698	-0.6695	-0.3960	58.27	<.0001
dist	1	1	0.2185	0.0585	0.1038	0.3332	13.93	0.0002
Dispersion		0	0.0000	0.0000	0.0000	0.0000		

The dispersion parameter is estimated to be 0, so that a Poisson regression model is actually fitted. Therefore, the results are identical with those in (2). This is not surprising as the deviance and Pearson Chi-square statistics in (2) indicate no overdispersion in the data.

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)

Solution: As the response variable has four categories and there is natural ordering in the categories (from no symptoms to serious symptoms), I would fit a proportional odds model for the response.

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

Solution: See "Assignment5.sas".

(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 the model results. (10 points)

Solution: SAS program to fit the proportional odds model is provided in "Assignment5.sas". The score test for the proportional odds assumption is given below:

Score Test for the Proportional Odds Assumption		
Chi-Square	DF	Pr > ChiSq
12.0745	8	0.1479

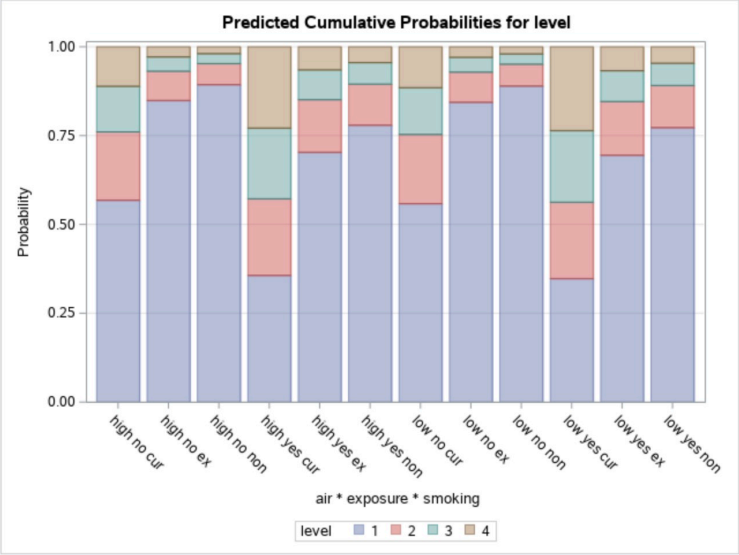
As the P-value = 0.1479 > 0.05, the proportional odds assumption is not rejected, i.e., the model is appropriate. The model results are:

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1	2.0884	0.1633	163.5861	<.0001
Intercept	2	1	2.9696	0.1693	307.7931	<.0001
Intercept	3	1	3.8938	0.1779	479.2836	<.0001
air	high	1	0.0393	0.0937	0.1758	0.6750
exposure	yes	1	-0.8648	0.0955	82.0603	<.0001
smoking	cur	1	-1.8527	0.1650	126.0383	<.0001
smoking	ex	1	-0.4000	0.2019	3.9267	0.0475

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
air high vs low	1.040	0.866	1.250
exposure yes vs no	0.421	0.349	0.508
smoking cur vs non	0.157	0.113	0.217
smoking ex vs non	0.670	0.451	0.996

The results indicate that:

- Subject's exposure to general air pollution (air) is not a significant variable related to the chronic respiratory disease status.
- Compared to subjects not exposed to pollution in their jobs, subjects exposed to pollution in their jobs tend to have more serious disease status.
- Compared to non-smokers, subjects who are ex-smokers and current smokers tend to have more serious disease status. Moreover, current smokers tend to have the worst disease status.



The predicted cumulative probabilities plot above shows that the high*yes*cur and low*yes*cur combinations have the lowest probability of having no symptoms (level I), i.e., group exposing to pollution in job and currently smoking has the worst disease status.