STA409

Answer to Assignment 5

12112627
李乐平

1. Solution.

(1). The log likelihood function of parameters is

$$l(\boldsymbol{\mu};\boldsymbol{y})=\sum_{i=1}^{n}y_i\ln\mu_i-\sum_{i=1}^{n}\mu_i-\sum_{i=1}^{n}\ln y_i!$$

It is obvious that the MLE. of $\boldsymbol{\mu}$ is $\mu_i=y_i$ for all i's, hence

$$l(\hat{\boldsymbol{\mu}}_S;\boldsymbol{y})=\sum_{i=1}^{n}y_i\ln y_i-\sum_{i=1}^{n}y_i-\sum_{i=1}^{n}\ln y_i!$$

Yet for the model of interest M, the MLE. is derived by

$$\begin{aligned}l(\hat{\boldsymbol{\mu}}_M;\boldsymbol{y})&=\sum_{i=1}^{n}y_i\ln\hat{\mu}_i-\sum_{i=1}^{n}\hat{\mu}_i-\sum_{i=1}^{n}\ln y_i!\\&=\sum_{i=1}^{n}y_i\ln\hat{y}_i-\sum_{i=1}^{n}\hat{y}_i-\sum_{i=1}^{n}\ln y_i!\end{aligned}$$

Thus, the deviance is given by

$$D(M)=2(l(\hat{\boldsymbol{\mu}}_S;\boldsymbol{y})-l(\hat{\boldsymbol{\mu}}_M;\boldsymbol{y}))=2(\sum_{i=1}^{n}y_i\ln\frac{y_i}{\hat{y}_i}-\sum_{i=1}^{n}(y_i-\hat{y}_i))$$

(2). We can know that

$$\mu_i=\exp(\beta_0+\sum_{k=1}^{p}\beta_k x_{ik})$$

Then the score statistic for β_0 is

$$\begin{aligned}U_0&=\frac{\partial l(\boldsymbol{\beta};\boldsymbol{y})}{\partial\beta_0}\\&=\sum_{i=1}^{n}y_i\frac{\partial\ln\mu_i}{\partial\beta_0}-\sum_{i=1}^{n}\frac{\partial\mu_i}{\partial\beta_0}\\&=\sum_{i=1}^{n}(y_i-\mu_i)\end{aligned}$$

(3). When applying the model M, we need to satisfy

$$U_0=\sum_{i=1}^{n}(y_i-\mu_i)=0$$

That is to say, the estimation must satisfies

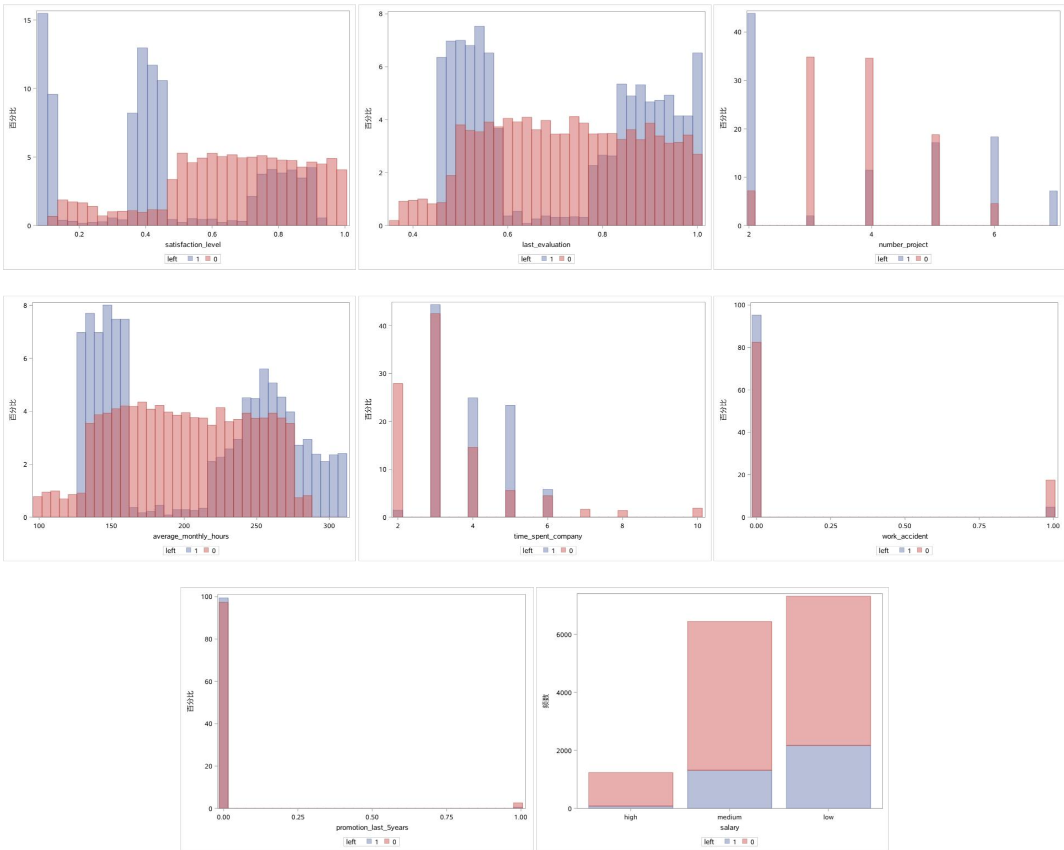
$$\sum_{i=1}^{n}\hat{y}_i=\sum_{i=1}^{n}\hat{\mu}_i=\sum_{i=1}^{n}y_i$$

Then the deviance is simplified to

$$\begin{aligned}D(M)&=2(\sum_{i=1}^{n}y_i\ln\frac{y_i}{\hat{y}_i}-\sum_{i=1}^{n}(y_i-\hat{y}_i))\\&=2\sum_{i=1}^{n}y_i\ln\frac{y_i}{\hat{y}_i}\end{aligned}$$

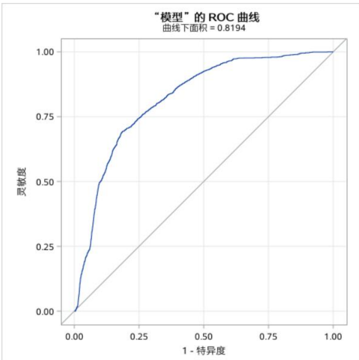
2. Solution.

(1). The bar plots below is drawn by the given data, showing the distribution of each variables grouped by "left". The distribution of resigned employees across continuous variables such as satisfaction_level, last_evaluation, and average_monthly_hours shows a multimodal pattern, whereas non-resigned employees exhibit an approximately uniform distribution.



(2). The AUC is 0.8194. That means a randomly chosen positive(left) sample has a probability of 0.8194 that its predicted value is higher than a random sampled negative one.

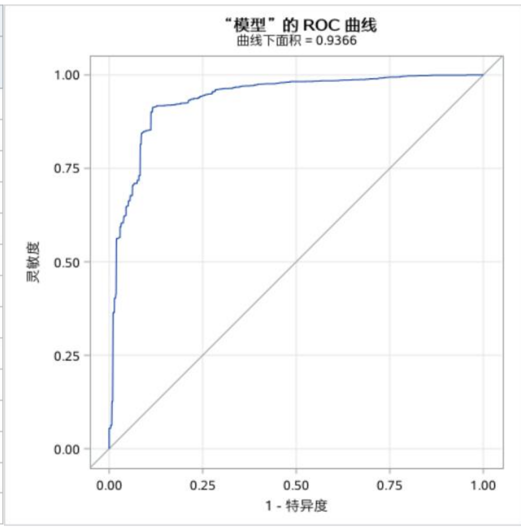
优比估计			
效应	点估计	95% Wald 置信限	
satisfaction_level	0.016	0.013	0.019
last_evaluation	2.068	1.545	2.767
number_project	0.730	0.700	0.761
average_monthly_hour	1.004	1.003	1.005
time_spent_company	1.299	1.260	1.338
work_accident 1 vs 0	0.215	0.181	0.256
promotion_last_5year 1 vs 0	0.227	0.138	0.375
salary high vs low	0.135	0.105	0.173
salary medium vs low	0.586	0.536	0.641



(3). By observing, we could take the following criterion to discretize the variables:

	satisfaction_level	last_evaluation	number_project	average_monthly_hour
low	<0.1	<0.6	<3	<160
medium	[0.1, 0.5)	[0.6, 0.82)	=3	[160, 260)
high	[0.5, 0.7)	>=0.82	>3	>=260
extra	>=0.7			

优比估计			
效应	点估计	95% Wald 置信限	
c_satisfaction_level hig vs low	<0.001	<0.001	>999.999
c_satisfaction_level med vs low	<0.001	<0.001	>999.999
c_satisfaction_level 彻 vs low	<0.001	<0.001	>999.999
c_last_evaluation hig vs low	3.219	2.743	3.777
c_last_evaluation med vs low	0.305	0.254	0.368
c_number_project hig vs low	0.107	0.090	0.127
c_number_project med vs low	0.010	0.008	0.014
c_average_monthly_ho hig vs low	2.409	2.005	2.895
c_average_monthly_ho med vs low	0.662	0.568	0.771
time_spent_company	1.435	1.379	1.493
work_accident 1 vs 0	0.225	0.183	0.277
promotion_last_5year 1 vs 0	0.257	0.146	0.451
salary high vs low	0.132	0.098	0.176
salary medium vs low	0.643	0.573	0.721



3. Solution.

(1).

CAR	AGE	DIST	y	n
1	1	0	65	317
1	1	1	2	20
1	4	1	36	316
1	3	1	4	40
1	2	1	5	33
1	3	0	52	486
1	2	0	65	476
1	4	0	310	3259
2	3	0	175	1355
2	2	0	159	1004
2	4	1	102	724
2	2	1	10	81
2	1	0	98	486
2	1	1	7	31
2	3	1	22	122
2	4	0	877	7660

(2). The table presents the goodness-of-fit criteria and maximum likelihood parameter estimates for a Poisson regression model. In terms of goodness of fit, we observe that the deviance and adjusted deviance are close to 1, indicating a good fit of the model to the data. Similarly, the Pearson chi-square and adjusted Pearson chi-square values being close to 1 suggest a good fit as well. Additionally, the small values of AIC and BIC further indicate that the model balances goodness of fit and model complexity well. Regarding the maximum likelihood parameter estimates, the intercept's estimate is -1.8102, implying that when all explanatory variables are zero, the logarithm of the expected value of the dependent variable is approximately 0.2. The coefficients for CAR, AGE, and DIST are positive, indicating that increases in these explanatory variables are associated with increases in the logarithm of the expected value of the dependent variable. Moreover, the significance tests for the coefficient estimates show that all explanatory variables are significantly different from zero, suggesting that these variables have a significant impact on the dependent variable.

最大似然参数估计的分析								
参数		自由度	估计	标准误差	Wald 95% 置信限		Wald 卡方	Pr > 卡方
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
尺度		0	1.0000	0.0000	1.0000	1.0000		

评估拟合优度的准则			
准则	自由度	值	值/自由度
偏差	24	23.7090	0.9879
调整后的偏差	24	23.7090	0.9879
Pearson 卡方	24	22.3393	0.9308
调整后的 Pearson X2	24	22.3393	0.9308
对数似然		14129.7072	
完全对数似然		-96.0346	
AIC（越小越好）		208.0693	
AICC（越小越好）		214.3302	
BIC（越小越好）		219.7952	

(3). The results provided are from a negative binomial regression model, assessing goodness of fit and maximum likelihood parameter estimates. In terms of goodness of fit, similar to the Poisson regression, the deviance, adjusted deviance, Pearson chi-square, and adjusted Pearson chi-square values are close to 1, indicating a good fit of the model to the data. The AIC and BIC values are relatively low, suggesting that the model balances goodness of fit and complexity effectively. However, there's a warning indicating potential convergence issues.

Moving to the maximum likelihood parameter estimates, the intercept's estimate is similar to the Poisson regression model, indicating that when all explanatory variables are zero, the logarithm of the expected value of the dependent variable is around 0.2. The coefficients for CAR, AGE, and DIST are also similar to the Poisson regression model, showing positive associations with the dependent variable. All of these coefficient estimates are statistically significant, indicating their impact on the dependent variable. Additionally, the dispersion parameter for the negative binomial regression model was estimated by maximum likelihood.

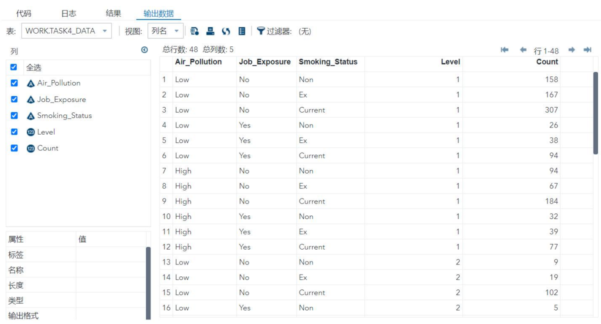
最大似然参数估计的分析								
参数		自由度	估计	标准误差	Wald 95% 置信限		Wald 卡方	Pr > 卡方
Intercept		1	-1.8102	0.0754	-1.9579	-1.6625	576.82	<.0001
CAR	2	1	0.1623	0.0678	0.0294	0.2952	5.73	0.0167
CAR	3	1	0.3935	0.0595	0.2770	0.5100	43.81	<.0001
CAR	4	1	0.5654	0.0733	0.4218	0.7090	59.54	<.0001
AGE	2	1	-0.1890	0.0829	-0.3514	-0.0266	5.20	0.0225
AGE	3	1	-0.3421	0.0824	-0.5037	-0.1806	17.23	<.0001
AGE	4	1	-0.5327	0.0699	-0.6698	-0.3957	58.06	<.0001
DIST	1	1	0.2185	0.0585	0.1038	0.3332	13.93	0.0002
离散度		1	0.0000	0.0065	.	.		

评估拟合优度的准则			
准则	自由度	值	值/自由度
偏差	24	23.7090	0.9879
调整后的偏差	24	23.7090	0.9879
Pearson 卡方	24	22.3386	0.9308
调整后的 Pearson X2	24	22.3386	0.9308
对数似然		14129.7072	
完全对数似然		-96.0346	
AIC（越小越好）		210.0693	
AICC（越小越好）		218.2511	
BIC（越小越好）		223.2609	

4. Solution.

(1). Since the response variable is an ordinal variable with 4 categories, a proportional odds model is suitable to be fitted.

(2).



	Air_Pollution	Job_Exposure	Smoking_Status	Level	Count
1	Low	No	Non	1	158
2	Low	No	Ex	1	167
3	Low	No	Current	1	307
4	Low	Yes	Non	1	26
5	Low	Yes	Ex	1	38
6	Low	Yes	Current	1	94
7	High	No	Non	1	94
8	High	No	Ex	1	67
9	High	No	Current	1	184
10	High	Yes	Non	1	32
11	High	Yes	Ex	1	39
12	High	Yes	Current	1	77
13	Low	No	Non	2	9
14	Low	No	Ex	2	19
15	Low	No	Current	2	102
16	Low	Yes	Non	2	5

(3). Since the p-value = 0.1479 > 0.05, the proportional odds assumption is not rejected.

比例优比假设的评分检验		
卡方	自由度	Pr > 卡方
12.0745	8	0.1479

The results suggest that general air pollution exposure is not significantly associated with chronic respiratory disease status. However, individuals exposed to pollution in their jobs tend to exhibit more severe disease statuses compared to those not exposed. Additionally, ex-smokers and current smokers generally have more severe disease statuses than non-smokers, with current smokers displaying the most severe conditions. The predicted cumulative probabilities plot indicates that individuals exposed to pollution in their jobs and currently smoking (high*yes*current combination) have the lowest probability of being symptom-free (level I), indicating the poorest disease status among the groups examined.

最大似然估计分析						
参数		自由度	估计	标准误差	Wald 卡方	Pr > 卡方
Intercept	1	1	1.2237	0.1748	48.9869	<.0001
Intercept	2	1	2.1049	0.1780	139.8649	<.0001
Intercept	3	1	3.0291	0.1841	270.5937	<.0001
Air_Pollution	High	1	0.0393	0.0937	0.1758	0.6750
Job_Exposure	No	1	0.8648	0.0955	82.0603	<.0001
Smoking_Status	Current	1	-1.8527	0.1650	126.0383	<.0001
Smoking_Status	Ex	1	-0.4000	0.2019	3.9267	0.0475

优比估计			
效应	点估计	95% Wald 置信限	
Air_Pollution High vs Low	1.040	0.866	1.250
Job_Exposure No vs Yes	2.374	1.969	2.863
Smoking_Status Current vs Non	0.157	0.113	0.217
Smoking_Status Ex vs Non	0.670	0.451	0.996