SAS. Assignment 5

毕圣杰 — 11910901.

Problem 1

1. Pmf of $NB(r,p)$: $f(y;r,p)=\binom{y+r-1}{y}(1-p)^r p^y=\binom{y+r-1}{y}\exp[y\log p+r\log(1-p)]$

$\Rightarrow$ let $h(y)=\binom{y+r-1}{y}$, $\theta=\log p$, $\phi=1$

then we have $p=e^\theta$, $A(\theta)=-r\log(1-p)=-r\log(1-e^\theta)$, $\to f(y;r,p)=h(y)\cdot\exp\left[\frac{y\theta-A(\theta)}{\phi}\right]$.

$\Rightarrow$ $NB(r,p)$ belongs to the exponential family.

2. Since $A(\theta)=-r\log(1-e^\theta)$, we have $E(Y)=A'(\theta)$, and $Var(Y)=\phi A''(\theta)=A''(\theta)$.

$\Rightarrow E(Y)=\frac{re^\theta}{1-e^\theta}=\frac{rp}{1-p}$ $\quad Var(Y)=\frac{re^\theta}{(1-e^\theta)^2}=\frac{rp}{(1-p)^2}$.

At the same time, $Var(Y)=\frac{rp}{(1-p)^2}=\frac{rp}{(1-p)}+\frac{rp^2}{(1-p)^2}=E(Y)+[E(Y)]^2/r$. □

3. Note that $A'(\theta)=\frac{re^\theta}{1-e^\theta}\to g(\mu)=(A')^{-1}(\mu)=\log\frac{\mu}{r+\mu}$.

$\Rightarrow$ the canonical link function: $g(\mu)=\log\frac{\mu}{r+\mu}$.

4. log-link function: $g_1(\mu)=\log\mu$.

canonical link function: $g(E(Y))=\log(\frac{rp}{1-p})=\log r+\text{logit}(p)=\beta_0+\beta_1x_1+\cdots+\beta_px_p$

1°. the logit function can map $p\in[0,1]$ to $\text{logit}(p)\in(-\infty,+\infty)$, while log function only map $p\in[0,1]$ to $\log p\in(-\infty,0]$.

2°. the logits function can easily compute odds and OR. which has better interpretation.

Problem 2

Note that $Y_1, \cdots, Y_n$ from the exponential distribution. $\theta_i \exp(-y_i\theta_i) = \exp(-\theta_i y_i + \log\theta_i)$

which belongs to canonical exponential family and $A(\theta) = -\log\theta$, $A'(\theta_i) = E(-y_i)$

$\Rightarrow \frac{1}{\theta_i} = E(y_i) = \mu_i$

→ likelihood function: $L(\underset{\sim}{\beta}, \underset{\sim}{y}) = \prod_{i=1}^{n} \theta_i \exp(-y_i\theta_i)$.

→ log-likelihood: $l(\underset{\sim}{\beta}, \underset{\sim}{y}) = \sum_{i=1}^{n} \log\theta_i - \sum_{i=1}^{n} y_i\theta_i \Rightarrow l(\underset{\sim}{\beta}) = \sum_{i=1}^{n} \log\frac{1}{\mu_i} - \sum_{i=1}^{n}\frac{y_i}{\mu_i}$

$\Rightarrow$ Saturated model: $l(\underset{\sim}{\hat{\mu}_s}) = \sum_{i=1}^{n} \log\frac{1}{y_i} - n$.

$\Rightarrow$ For model M: $f(y_i;\theta) = \theta\exp(-y_i\theta_i) \rightarrow l(\hat{\beta}_M) = \sum_{i=1}^{n}\log\theta - \sum_{i=1}^{n} y_i\theta = \sum_{i=1}^{n}(\log\theta - y_i\theta)$.

Note that $\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n}(\frac{1}{\theta} - y_i)\underset{\sim}{x_i} = 0$. $\Rightarrow \sum_{i=1}^{n}(\frac{1}{\theta} - y_i) = 0 \rightarrow \theta = \frac{1}{\bar{y}}$

$\Rightarrow l(\hat{\beta}_M) = \sum_{i=1}^{n}(\log\frac{1}{\bar{y}} - \frac{y_i}{\bar{y}}) = -n\log\bar{y} - n$.

$\Rightarrow D = 2(l(\mu_s) - l(\hat{\beta}_M)) = 2\left(\sum_{i=1}^{n}(\log\frac{1}{y_i} + \log\bar{y})\right) = 2\sum_{i=1}^{n}\log\frac{\bar{y}}{y_i} = -2\sum_{i=1}^{n}\log\frac{y_i}{\bar{y}}$

By Jensen's inequality,

We have $\frac{\sum_{i=1}^{n}\log\frac{y_i}{\bar{y}}}{n} \leq \log\left(\frac{\sum_{i=1}^{n}\frac{y_i}{\bar{y}}}{n}\right) \Rightarrow \sum_{i=1}^{n}\log\frac{y_i}{\bar{y}} \leq 0$.

$\Rightarrow D = -2\sum_{i=1}^{n}\log\frac{y_i}{\bar{y}} \geq 0$

$\Rightarrow$ Hence, the deviance is always non-negative.

Problem 3.

(2). 1° The parameter estimation of the logistic regression model on the prob. of drinking alcohol frequently are given below:

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	-2.1140	0.2715	60.6485	<.0001
EI	I	1	-0.5550	0.2170	6.5422	0.0105
SN	S	1	-0.4292	0.2340	3.3641	0.0666
TF	T	1	0.6873	0.2206	9.7067	0.0018
JP	P	1	0.2022	0.2266	0.7967	0.3721

2°

→ the odds ratio estimate are given below:

Odds Ratio Estimates

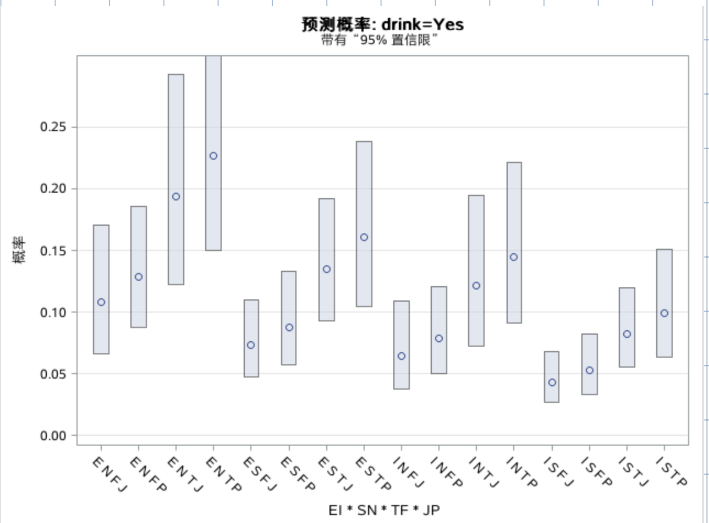
Effect	Point Estimate	95% Wald Confidence Limits	
EI E vs I	1.742	1.139	2.665
SN S vs N	0.651	0.412	1.030
TF T vs F	1.988	1.290	3.064
JP J vs P	0.817	0.524	1.274

⇒ explanation:

- People with Extroversion/Introversion scale is E has 1.742 times of drinking alcohol frequently compared to people with I.
- People with Sensing/iNtuitive scale is S has 0.651 times of drinking alcohol frequently compared to people with N
- People with Thinking/Feeling scale is T has 1.988 times of drinking alcohol frequently compared to people with F.
- People with Judging/Perceiving scale is J has 0.817 times of drinking alcohol

frequently compared to people with P.

3°. The predictived probabilities for the 16 personality type is given below:



The person with ENTP has the highest predicted probability of drinking alcohol frequently

(3). The odds ratio estimates are given below:

优比估计和 Wald 置信区间			
优比	估计	95% 置信限	
SN S vs N at TF=F	0.420	0.231	0.761
SN S vs N at TF=T	1.032	0.502	2.121

People with Thinking/Feeling scale is T. Sensing people had 0.42 times the odds of drinking alcohol frequently compared to iNtuitive people.

People with Thinking/Feeling scale is F. Sensing people had 1.032 times the odds of drinking alcohol frequently compared to iNtuitive people.

Problem 4.

(1). From the pie figure 1, there are 23.81% employees ended up leaving the company.

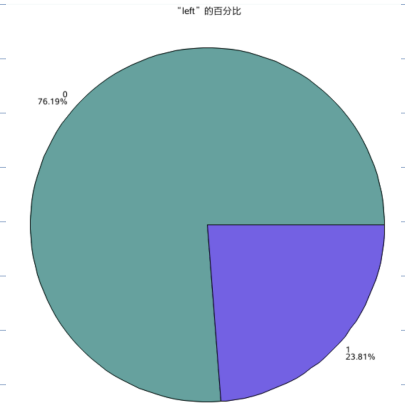


figure 1

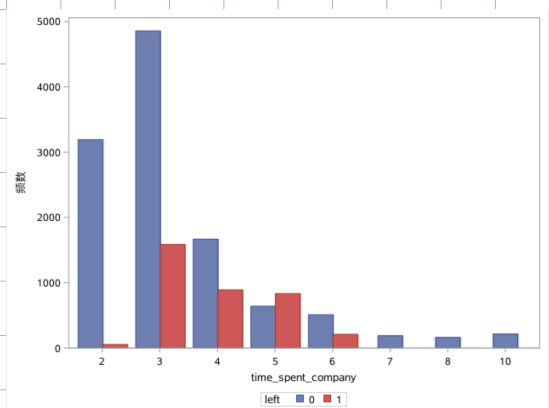


figure 2.

1° From figure 2, people who have been with the company for more than 7 years have not left. 2° The number of people who just complete 2 years in the company have relatively small proportion to leave

For those who have been with the company for 3~5 years, the proportion of leaving the company is higher.

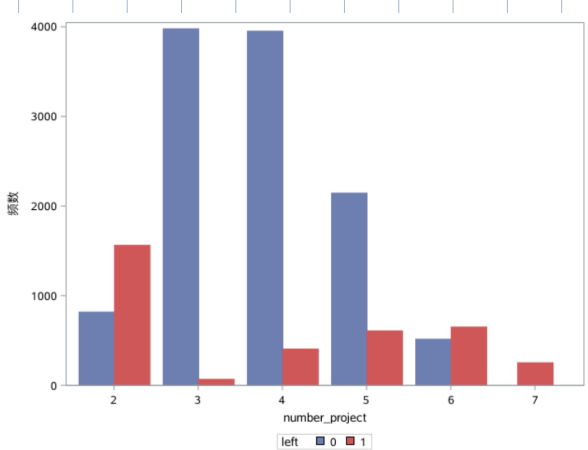


figure 3.

Employees with the fewest and most projects had a higher percentage of leaving the company, and people with 3 projects have the lowest percentage of leaving the company

The highest percentage of people leaving the company is among the high-paying group, while the highest percentage is among the middle-paying group

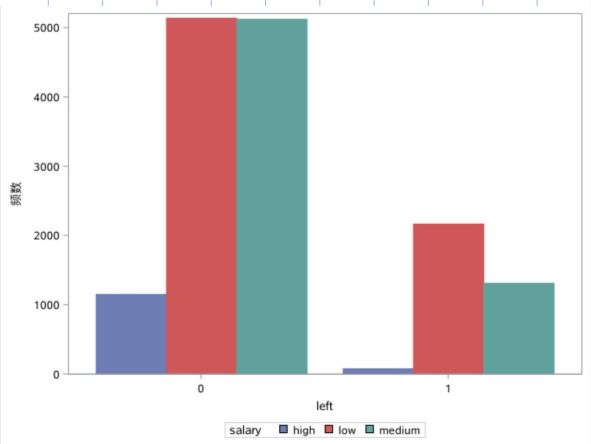


figure4

People who leave the company tend to be less satisfied with the company than those who do not leave the company, and the variance of the degree of satisfaction is also larger

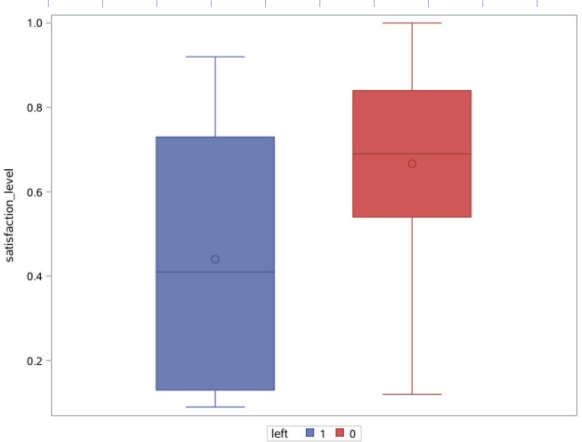


figure 5

(2). The odds ratio estimates are given below:

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

⇒ explanations:

1°. When other explanatory variables are fixed, if the satisfaction level / last evaluation / number of projects worked on / average monthly hours / time spent in the company increase by 1 unit, the odds of people leaving the company changes by multipling 0.016 / 2.068 / 0.73 / 1.004 / 1.299.

2°. The people got a work accident in the last 2 years has 0.215 times the odds of leaving the company compared to people who don't.

3°. The people got a promotion in the last 5 years has 0.227 times the odds of leaving the company compared to people who don't.

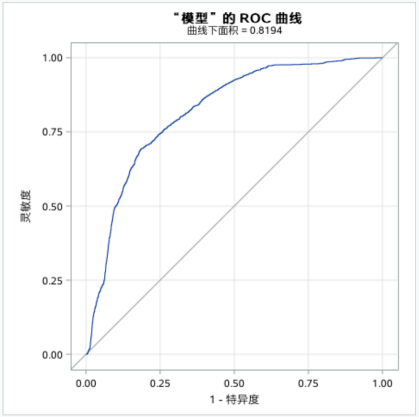
4°. The people with high salary has 0.135 times the odds of leaving the company compared to people with low salary.

The people with median salary has 0.586 times the odds of leaving the company compared to people with low salary.

⇒ then the Roc curve is given below:

from the table, the AUC is 0.8194, indicating the probability that a randomly chosen positive samples has a higher predicted value than a randomly

预测概率和观测响应的关联			
一致部分所占百分比	81.9	Somers D	0.639
不一致部分所占百分比	18.1	Gamma	0.639
结值百分比	0.0	Tau-a	0.232
对	40809388	c	0.819



chosen negative one is 0.8194

(3). We take the transformation to the following variables:

$$\text{Satisfaction\_level} = \begin{cases} \text{yes} & \text{if satisfaction level} \le 0.5 \\ \text{no} & \text{otherwise} \end{cases}$$

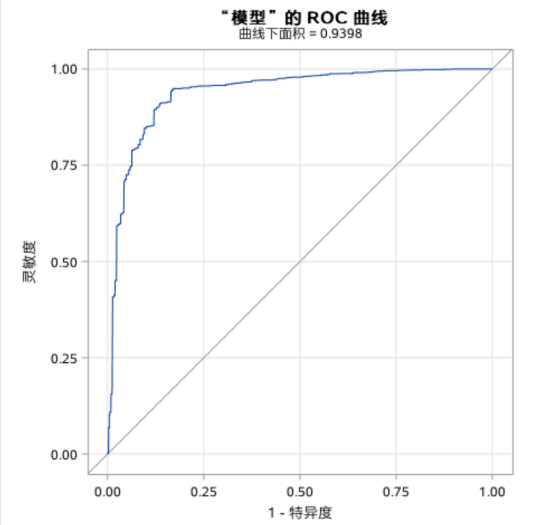
$$\text{number project} = \begin{cases} \text{low} & \text{if number project} = 2 \\ \text{median} & \text{if number project} \in [3,4] \\ \text{high} & \text{otherwise} \end{cases}$$

$$\text{last evaluation} = \begin{cases} \text{low} & \text{if last evaluation} \le 0.6 \\ \text{median} & \text{if } 0.6 < \text{last evaluation} \le 0.8 \\ \text{high} & \text{otherwise} \end{cases}$$

$$\text{average monthly hours} = \begin{cases} \text{low:} & \text{if average monthly hours} \le 165 \\ \text{median:} & \text{if } 165 < \text{average monthly hours} \le 215 \\ \text{high:} & \text{if } 215 < \text{average monthly hours} \le 275 \\ \text{very high:} & \text{otherwise} \end{cases}$$

$$\text{time spend Company} = \begin{cases} \text{short} & \text{if time spend Company} = 2 \\ \text{median} & \text{if } 3 \le \text{time spend Company} \le 6 \\ \text{long} & \text{otherwise} \end{cases}$$

the new AUC curve is given below: with ROC = 0.9398



Problem 5.

(2). The estimated parameters of Poisson regression is given below:

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	-0.1143	0.0588	-0.2295	0.0010	3.78	0.0520
Duration	10-14	1	1.3702	0.0511	1.2700	1.4704	718.35	<.0001
Duration	15-19	1	1.6133	0.0513	1.5127	1.7139	987.91	<.0001
Duration	20-24	1	1.7834	0.0514	1.6826	1.8841	1203.24	<.0001
Duration	5-9	1	0.9977	0.0528	0.8943	1.1012	357.65	<.0001
Education	LP	1	0.0070	0.0276	-0.0472	0.0612	0.06	0.7999
Education	SH	1	-0.3128	0.0567	-0.4239	-0.2017	30.45	<.0001
Education	UP	1	-0.1108	0.0346	-0.1787	-0.0429	10.23	0.0014
place	Rural	1	0.1564	0.0347	0.0885	0.2244	20.37	<.0001
place	Urban	1	0.1251	0.0396	0.0475	0.2027	9.97	0.0016
尺度		0	1.0000	0.0000	1.0000	1.0000		

From the above result:

1°. # of children on average ever born of Indian race increases with the increasing of education level.

2°. #. of children on average ever born of Indian race decrease with the prosperity of the place of residence increase.

3°. there not exist significant difference on #. of children on average ever born of Indian race for the people with different educational level.

(3). The estimation of parameters are given below:

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	-0.1143	0.0593	-0.2304	0.0019	3.72	0.0539
Duration	10-14	1	1.3702	0.0512	1.2699	1.4705	717.25	<.0001
Duration	15-19	1	1.6133	0.0515	1.5122	1.7143	979.42	<.0001
Duration	20-24	1	1.7834	0.0542	1.6772	1.8895	1084.23	<.0001
Duration	5-9	1	0.9977	0.0533	0.8933	1.1021	350.88	<.0001
Education	LP	1	0.0070	0.0280	-0.0479	0.0620	0.06	0.8027
Education	SH	1	-0.3128	0.0574	-0.4252	-0.2004	29.74	<.0001
Education	UP	1	-0.1108	0.0351	-0.1795	-0.0421	9.99	0.0016
place	Rural	1	0.1564	0.0348	0.0882	0.2247	20.17	<.0001
place	Urban	1	0.1251	0.0399	0.0470	0.2032	9.85	0.0017
离散度		1	0.0000	0.0012	.	.		

We find that the result of two regression models are exactly same. And the dispersion ϕ of negative binomial regression model is 1. → We donot need to adjust the overdispersion.