의학통계학 과제 #2

영어영문학과 201610010 김유민

1.

```
DATA jail;
INPUT live $ arrest $ jail $ count @@;
CARDS;
n y y 42 n y n 109 n n y 17 n n n 75
y y y 33 y y n 175 y n y 53 y n n 359
;

PROC LOGISTIC DESCENDING;
FREQ count;
CLASS live arrest;
MODEL jail = live arrest / SCALE = NONE AGGREGATE;
RUN;
```

Sum of Frequencies Read	138
Sum of Frequencies Used	138

Response Profile

Ordered Value	Binary Outcome	Total Frequency
1	Event	9
2	Nonevent	129

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	11.9564	14	0.8540	0.6098
Pearson	13.5722	14	0.9694	0.4820

Number of unique profiles: 16

Model Fit Statistics

-2 Log L	66.540	60.396	31.647

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	6.1440	1	0.0132
Score	6.7696	1	0.0093
Wald	6.0435	1	0.0140

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	5.0850	3.0525	2.7751	0.0957
temp	1	-0.1156	0.0470	6.0435	0.0140

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
temp	0.891	0.812	0.977

Association of Predicted Probabilities and Observed Responses

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	783.413	771.323
SC	788.173	785.604
-2 Log L	781.413	765.323

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	16.0899	2	0.0003
Score	16.8252	2	0.0002
Wald	16.3722	2	0.0003

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
live	1	9.0509	0.0026
arrest	1	3.3127	0.0687

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.4828	0.0951	243.2458	<.0001
live	n	1	0.2960	0.0984	9.0509	0.0026
arrest	n	1	-0.1735	0.0953	3.3127	0.0687

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
live n vs y	1.808	1.229	2.658
arrest n vs y	0.707	0.486	1.027

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	44.7	Somers' D	0.188
Percent Discordant	25.9	Gamma	0.266
Percent Tied	29.4	Tau-a	0.053
Pairs	104110	c	0.594

(1)

먼저, 유의수준 $\alpha = 0.05$에서 로짓 모형의 적합도 검정을 실시한다. 통계적 가설은 다음과 같다.

H_0 : 로짓 모형이 적합하다.

H_1 : 로짓 모형이 적합하지 않다.

위 SAS 결과 중 'Deviance and Pearson Goodness-of-Fit Statistics'를 보면

우도비 카이제곱 통계량의 p값 0.4762 > 유의수준 0.05

피어슨 카이제곱 통계량의 p값 0.4784 > 유의수준 0.05

이므로 귀무가설을 기각할 수 없다. 따라서 유의수준 0.05에서 로짓 모형이 적합함을 알 수 있다.

다음으로 모든 회귀 계수 추정치에 대한 가설검정결과를 확인한다. 통계적 가설은 다음과 같다.

$$H_0 : \beta_1 = \beta_2 = 0$$

$$H_1 : \text{적어도 하나의 } \beta_i (i = 1,2) \text{는 0이 아니다.}$$

위 SAS 결과 중 'Testing Global Null Hypothesis: BETA=0'를 보면 우도비, 스코어, 왈드 통계량 모두 p값이 0.05보다 작으므로 유의수준 0.05에서 귀무가설을 기각한다. 따라서 추정치들의 유의미함을 알 수 있다.

다음으로 'Type 3 Analysis of Effects' 표를 보고 모형의 변수의 유의성을 확인한다.
이 때, 1행의 가설은 $H_0 : \beta_1 = 0$에 대해 $H_1 : \beta_1 \neq 0$이고 2행의 가설은 $H_0 : \beta_2 = 0$에 대해 $H_1 : \beta_2 \neq 0$ 이다.
거주 여부 변수(β_1)의 유의성 검정에서 p값은 0.0026로 유의수준 $\alpha = 0.05$보다 작으므로 귀무가설을 기각한다.
체포 경력 변수(β_2)의 유의성 검정에서 p값은 0.0687로 유의수준 $\alpha = 0.05$보다 크므로 귀무가설을 기각하지 못한다.
즉, 거주 여부에 대한 변수는 구속확률에 유의한 영향을 끼치고, 체포경력은 구속확률에 유의한 영향을 끼치지 않는다.

마지막으로 'Analysis of Maximum Likelihood Estimates' 표를 보고 로짓 모형을 세우면

$$\log(\frac{p}{1-p}) = -1.4828 + 0.296live - 0.1735arrest$$

인데, 앞에서 체포경력은 구속확률에 유의한 영향을 끼치지 않는다($\beta_2 = 0$)고 하였으므로 최종 모형은

$$\log(\frac{p}{1-p}) = -1.4828 + 0.296live \text{ 이다.}$$

단, 여기서 거주민이 아니면 (No) live =1 이고 거주민이면(Yes) live = -1 이다.

(2)
위 SAS 결과 중 'Odds Ratio Estimates' 표를 통해 절도 피의자가 구속될 오즈를 살펴본다.
절도 피의자가 구속될 오즈를 거주민 여부에 따라 보면 체포경력을 고정시킨 상태에서 '거주민이 아니다(No)'와 '거주민이다(Yes)'의 오즈비는 1.808이고 신뢰구간(1.229,2.658)에 1이 포함되지 않으므로 오즈비가 유의미하다고 할 수 있다. 즉, 거주민이 아닐 때 절도 피의자가 구속될 확률의 오즈는 거주민일 때의 180.8%이다.
절도 피의자가 구속될 오즈를 체포경력에 따라 보면 거주민 여부를 고정시킨 상태에서 체포경력이 없다(No)'와 체포경력이 있다(Yes)'의 오즈비는 0.707이다. 즉, 체포경력이 없을 때 구속확률의 오즈는 체포경력이 있을 때의 70.7%이다. 그러나 신뢰구간(0.486,1.027)에 1이 포함되므로 오즈비가 유의미하지 않다고 할 수 있다. 따라서 체포경력과 구속여부는 연관이 없다.

2.

```
DATA lungcancer;
INPUT cancer $ smoking $ count @@;
CARDS;
1 1 9
```

```
1 0 5
0 1 46
0 0 40
;
RUN;

PROC LOGISTIC DATA = lungcancer DESCENDING;
FREQ count;
CLASS smoking cancer;
MODEL cancer = smoking / SCALE = NONE AGGREGATE;
RUN;
```

Number of Observations Read	4
Number of Observations Used	4
Sum of Frequencies Read	100
Sum of Frequencies Used	100

Response Profile

Ordered Value	cancer	Total Frequency
1	1	14
2	0	86

Probability modeled is cancer='1'.

Class Level Information

Class	Value	Design Variables
smoking	0	1
	1	-1

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Wald	0.5606	1	0.4540

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
smoking	1	0.5606	0.4540

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.8553	0.2991	38.4801	<.0001
smoking	0	1	-0.2239	0.2991	0.5606	0.4540

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
smoking 0 vs 1	0.639	0.198	2.064

Association of Predicted Probabilities and Observed Responses

Percent Concordant	29.9	Somers' D	0.108
Percent Discordant	19.1	Gamma	0.220

(1)

먼저, 표 Response Profile 정보를 먼저 확인한다. Ordered Value에서 반응변수 'cancer'의 수준 중에서 '1'(Yes) 가 1의 값을 갖는다는 것을 확인할 수 있다. 또한 표 Class Level

Information에서 변수 'smoking'의 수준은 '0'과 '1'로 코딩되었지만, 모형 적합과정에서는 1과 -1로 인식됨을 알 수 있다.

다음으로 'Type 3 Analysis of Effects' 표를 보고 모형의 변수의 유의성을 확인한다.
이 때, 가설은 $H_0 : \beta_1 = 0$에 대해 $H_1 : \beta_1 \neq 0$이다.
흡연 여부 변수(β_1)의 유의성 검정에서 p값은 0.454로 유의수준 $\alpha = 0.05$보다 크므로 귀무가설을 기각할만한 충분한 근거가 없다.
즉, 흡연여부는 폐암 발생률에 유의한 영향을 끼치지 않는다.

마지막으로 'Analysis of Maximum Likelihood Estimates' 표를 보고 로짓 모형을 세우면

$$\log(\frac{p}{1-p}) = -1.8553 - 0.2239 smoking$$

인데, 앞에서 흡연 여부 변수은 폐암 발생률에 유의한 영향을 끼치지 않는다($\beta_1 = 0$)고 하였으므로 이 모형에서 smoking은 모형에 대해 유의미하다고 할 수 없다.
단, 여기서 비흡연자라면(0) smoking =1 이고 흡연자라면(1) smoking = -1 이다.

(2)
표 'Odds Ratio Estimates'를 보면, 오즈비는 0.639이다. 흡연 여부 변수인 smoking이 '0'(No)일 때의 오즈는 $e^{0.2239}$이고 smoking이 '1'(Yes)일 때의 오즈는 $e^{-0.2239}$이므로 smoking변수 '0'(No)와 '1'(Yes)의 오즈비는 $e^{0.4478} = 0.639$인 것을 도출할 수 있다.

(3)
위에서 구한 회귀계수의 추정치를 통해 흡연 여부 별로 로짓 모형식을 쓰면 다음과 같다.

$$\text{비흡연 No} : \frac{p_{x(n)}}{1-p_{x(n)}} = \exp(-1.8553 - 0.2239)$$

$$\text{흡연 Yes} : \frac{p_{x(y)}}{1-p_{x(y)}} = \exp(-1.8553 + 0.2239)$$

이를 통해 smoking 변수 '0'와 '1'의 오즈비는 다음과 같이 나타난다.

$$\frac{\frac{p_{x(n)}}{1-p_{x(n)}}}{\frac{p_{x(y)}}{1-p_{x(y)}}} = \exp(-0.2239 - 0.2239) = \exp(-0.4478) = 0.639$$

따라서, 흡연 여부가 '0'(No)일 경우 '1'(Yes)인 경우보다 폐암이 발병할 오즈가 0.6배나 낮은 것을 알 수 있다.

3.

```
DATA car;
INPUT x1 x2 y;
CARDS;
45 2 0
40 4 1
```

```
60 3 0
50 2 0
55 2 0
50 5 1
35 7 1
65 10 1
53 2 0
48 1 0
37 5 1
31 7 1
40 4 0
75 8 1
43 9 1
49 2 0
37.5 4 1
71 11 1
34 5 0
27 6 1
;
RUN;

PROC LOGISTIC DESCENDING data =car outmodel=predmodel noprint;
MODEL y = x1 x2 / SCALE = NONE AGGREGATE;
RUN;
```

Number of Observations Read	20
Number of Observations Used	20

Response Profile

Ordered Value	y	Total Frequency
1	1	11
2	0	9

Probability modeled is y=1.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	6.4822	16	0.4051	0.9820
Pearson	6.5858	16	0.4116	0.9804

Number of unique profiles: 19

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	18.2707	2	0.0001
Score	12.7644	2	0.0017
Wald	3.9165	2	0.1411

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-8.0025	9.1589	0.7634	0.3823
x1	1	0.0117	0.1344	0.0075	0.9309
x2	1	1.8225	1.1126	2.6832	0.1014

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
x1	1.012	0.777	1.317
x2	6.187	0.699	54.774

Association of Predicted Probabilities and

(1)

먼저, 표 Response Profile 정보를 먼저 확인한다. Ordered Value에서 반응변수 'y'의 값이 순서대로 나열되어 있는 것을 알 수 있다. 하단의 'Probability modeled is y=1.'를 통해서 성공확률은 y변수가 1의 값을 가질 때의 확률이라는 것을 알 수 있다.

다음으로 유의수준 $\alpha = 0.05$에서 로짓 모형의 적합도 검정을 실시한다. 통계적 가설은 다음과 같다.

H_0 : 로짓 모형이 적합하다.

H_1 : 로짓 모형이 적합하지 않다.

위 SAS 결과 중 'Deviance and Pearson Goodness-of-Fit Statistics'를 보면

우도비 카이제곱 통계량의 p값 0.9820 > 유의수준 0.05

피어슨 카이제곱 통계량의 p값 0.9804 > 유의수준 0.05

이므로 귀무가설을 기각할 수 없다. 따라서 유의수준 0.05에서 로짓 모형이 적합함을 알 수 있다.

마지막으로 'Analysis of Maximum Likelihood Estimates' 표를 보고 로짓 모형을 세우면

$$\log(\frac{p}{1-p}) = -8.0025 + 0.0117x1 + 1.8225x2$$

인데, 이 때 intercept, x1, x2의 p값이 각각 0.3823, 0.9309, 0.1014 로 모두 0.05보다 크므로 이 모형에 대한 각각의 추정치는 유의하다고 할 수 없다.

(2)

위에서 구한 회귀계수의 추정치 중 가구 소득(x1)이 고정되었다고 가정하면 차량 보유기간(x2)이 1년 증가했을 때 새 차를 구입할 확률을 알 수 있다.

a. 차량보유기간이 (r+1)년일 때 : $\log\frac{p(x1, r+1)}{1-p(x1, r+1)} = -8.0025 + 0.0117x1 + 1.8225(r+1)$

b. 차량보유기간이 r년일 때 : $\log\frac{p(x1, r)}{1-p(x1, r)} = -8.0025 + 0.0117x1 + 1.8225r$

위 a와 b의 차를 구하면

$$\frac{\frac{p(x1,r)}{1-p(x1,r)}}{\frac{p(x1,r+1)}{1-p(x1,r+1)}} = \exp(1.8225) \fallingdotseq 6.187$$

따라서 차량 보유기간이 1년 증가하면 새 차를 구입할 확률이 약 6배 높다는 것을 알 수 있다.

한편 위에서 구한 회귀계수의 추정치 중 차량보유기간(x2)이 고정되었다고 가정하면 가구 소득이 백만원 증가했을 때 새 차를 구입할 확률을 알 수 있다.

a. 가구 소득이 (r+1)*백만원일 때 :

$$\log\frac{p(r+1,x2)}{1-p(r+1,x2)} = -8.0025 + 0.0117(r+1) + 1.8225x2$$

b. 가구 소득이 r*백만원일 때 : $\log\frac{p(r,x2)}{1-p(r,x2)} = -8.0025 + 0.0117r + 1.8225x2$

위 a와 b의 차를 구하면

$$\frac{\frac{p(x1,r)}{1-p(x1,r)}}{\frac{p(x1,r+1)}{1-p(x1,r+1)}} = \exp(0.0117) \fallingdotseq 1.012$$

따라서 가구 소득이 백만원 증가하면 새 차를 구입할 확률이 약 1.012배 높다는 것을 알 수 있으며, 이는 큰 차이가 없음을 알 수 있다.

(3)

```
DATA new;
INPUT x1 x2 ;
CARDS;
44 2.5
;
run;

PROC LOGISTIC inmodel=predmodel  ;
    score data=new out=newprob;
RUN;

proc print data=newprob; run;
```

OBS	x1	x2	I_y	P_1	P_0
1	44	2.5	0	0.050527	0.94947

위 SAS 결과에 따르면 가구 1년 소득이 44백만원이고 가장 오래된 차의 보유기간이 2.5년인 경우 구입할 확률이 5%, 구입하지 않을 확률이 약 95%이기 때문에 새 차를 구입하지 않을 확률이 높다.

4.

```
Data oring;
input trial failed temp ;
cards;
6 0 66
6 1 70
6 0 69
6 0 68
6 0 67
6 0 72
6 0 73
6 0 70
6 1 57
6 1 63
6 1 70
6 0 78
6 0 67
6 2 53
6 0 67
6 0 75
6 0 70
6 0 81
6 0 76
6 0 79
6 2 75
6 0 76
6 1 58
;
run;

proc logistic data=oring;
  model failed/trial = temp ;
run;
```

Sum of Frequencies Read	138
Sum of Frequencies Used	138

Response Profile

Ordered Value	Binary Outcome	Total Frequency
1	Event	9
2	Nonevent	129

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	11.9564	14	0.8540	0.6098
Pearson	13.5722	14	0.9694	0.4820

Number of unique profiles: 16

-2 Log L	66.540	60.396	31.647

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	6.1440	1	0.0132
Score	6.7696	1	0.0093
Wald	6.0435	1	0.0140

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	5.0850	3.0525	2.7751	0.0957
temp	1	-0.1156	0.0470	6.0435	0.0140

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
temp	0.891	0.812	0.977

Association of Predicted Probabilities and Observed Responses

(1)

먼저, 표 Response Profile 정보를 먼저 확인한다. Ordered Value에서 반응변수 'failed/trial'의 수준 중에서 'Event', 즉 파손한 사건이 1의 값을 갖는다는 것을 확인할 수 있다.

다음으로 유의수준 $\alpha = 0.05$에서 로짓 모형의 적합도 검정을 실시한다. 통계적 가설은 다음과 같다.

H_0 : 로짓 모형이 적합하다.

H_1 : 로짓 모형이 적합하지 않다.

위 SAS 결과 중 'Deviance and Pearson Goodness-of-Fit Statistics'를 보면

우도비 카이제곱 통계량의 p값 0.6098 > 유의수준 0.05

피어슨 카이제곱 통계량의 p값 0.4820 > 유의수준 0.05

이므로 귀무가설을 기각할 수 없다. 따라서 유의수준 0.05에서 로짓 모형이 적합함을 알 수 있다.

다음으로 모든 회귀 계수 추정치에 대한 가설검정결과를 확인한다. 통계적 가설은 다음과 같

다.

$$H_0 : \beta_1 = \beta_2 = 0$$

$$H_1 : \text{적어도 하나의 } \beta_i (i = 1,2) \text{는 0이 아니다.}$$

위 SAS 결과 중 'Testing Global Null Hypothesis: BETA=0'를 보면 우도비, 스코어, 왈드 통계량 모두 p값이 0.05보다 작으므로 유의수준 0.05에서 귀무가설을 기각한다. 따라서 추정치들의 유의미함을 알 수 있다.

마지막으로 'Analysis of Maximum Likelihood Estimates' 표를 보고 로짓 모형을 세우면

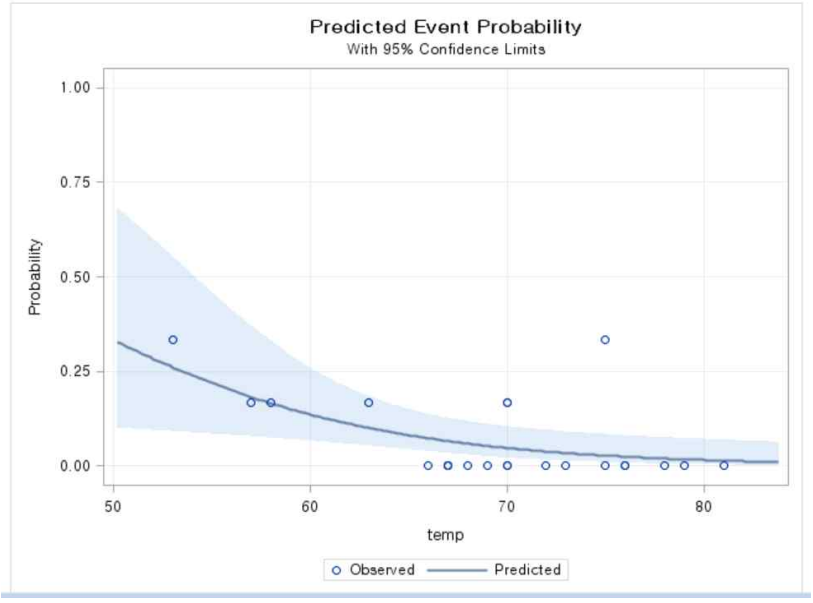
$$\log(\frac{p}{1-p}) = 5.0850 - 0.1156 temp$$

인데, 이 때 intercept의 p값이 0.0957로 0.05보다 크므로 이 모형에 대한 추정치 α는 유의하지 않다.

(2)

```
ods graphics on;
PROC LOGISTIC data=oring  plots(only)=(effect);;
MODEL failed/trial = temp / SCALE=NONE AGGREGATE;
output out=pdata;
RUN;
```

위 SAS 코드를 통해 온도의 변화에 따른 파손확률의 변화를 graph로 그리면 다음과 같다.



위의 그래프를 통해 온도가 높아질수록 파손확률이 낮아진다는 것을 알 수 있다.

(3)

```
PROC LOGISTIC DESCENDING data =oring outmodel=predmodel noprint;
MODEL failed/trial = temp / SCALE = NONE AGGREGATE;
RUN;
```

```
DATA new;
INPUT temp ;
CARDS;
50
40
30
;
run;

PROC LOGISTIC inmodel=predmodel  ;
    score data=new out=newprob;
RUN;

proc print data=newprob; run;
```

OBS	temp	_INTO_	P_Event	P_Nonevent
1	50	Nonevent	0.33290	0.66710
2	40	Event	0.61323	0.38677
3	30	Event	0.83437	0.16563

위 SAS 프로그램 결과에 따르면 온도가 50도일 경우 파손확률 추정값은 약 33.29%이고, 온도가 40도일 경우는 약 61.323%, 온도가 30도일 경우에는 파손확률 추정값이 약 83.437%이다.