Chapter 9 로지스틱 회귀분석

이제까지 회귀분석은 종속변수가 연속형인(metric) 경우 살펴보았으나 이 장에서는 종속변수가 이진(binary: 가질 수 있는 값이 실패/성공, 정품/불량 등과 같이 가질 수 있는 값이 2개인 경우)인 경우 사용되는 회귀분석 방법인 로지스틱 회귀분석(Logistic regression)을 살펴볼 것이다. 로지스틱 회귀분석의 설명변수는 측정형과 분류형(지시 변수)가 가능하지만 회귀분석처럼 지시 변수가 너무 많으면 모형이 복잡해지고 해석이 복잡해진다.

로지스틱 회귀분석에서 종속변수 값은 0, 1(사건: 성공, 불량)로 입력된다. 로지스틱 회귀분석은 이진형 반응변수 뿐 아니라 반응변수가 순서형(ordinal) 분류형인 경우 사용할 수 있다. 예를 들면 종속 변수가 고객의 신용도이고 이 변수가 (상, 중, 하) 분류되어 있는 경우 사용할 수 있다. 종속변수의 수준이 3개 이상인 경우 LOGISTIC 모형을 사용하는 것이 아니라 CATMOD를 사용해야 한다고 언급한 책이 있다. 그러나 CATMD는 CATegorical data MODeling의 약어로 분류변수 자료 모형화이며, LOGISTIC 모형은 CATMOD 기법의 한 부분이다.

9.1 로지스틱 모형

9.1.1 일반 선형 회귀 모형 $y_i = f(\underline{x}) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + .. + \beta_p x_{pi} + e_i$, $e_i \sim iidN(0, \sigma^2)$

로지스틱 회귀모형의 종속 변수는 0과1 두 값만 가지므로(더 이상 정규분포를 따르지 않는다) 결정계수(R^2)가 매우 낮고(이산형 변수의 문제점, 설문 분석의 Likert 척도 문항도 같은 문제) F-검정이나 t-검정을 사용하여 모형, 회귀 계수 유의성 검정을 하는데 문제가 있다. 가장 큰 문제는 종속 변수 y_i가 이진형인 경우(자료가 0, 1만 존재) OLS에 의한 계수 추정은 무의미 하여 음의 값이 예측되거나 부호 자체가 달라지게 된다. 또한 이진형 변수의 특성상 이분산 가능성이 매우 높다.

9.1.2 ODDS

$ODDs = \frac{p}{1-p}$ 로 정의되며 p 임의의 사건이 발생할(성공) 확률로 이것은 도박의 기준이 된다. 한국이 2002년 16강에 들어갈 확률 0.1이면 1/9이 Odds이다. 즉 한국 승리에 1$을 걸은 사람은 한국이 이길 경우 9$을 상금으로 받게 된다. 브라질이 2002년 16강에 들어갈 확률 0.8이면 4가 Odds이다. 그러므로 4$을 걸면 1$을 상금으로 받게 된다.

9.1.3 ODDS transformation $p^* = \frac{p}{1-p}$

종속변수를 $y_i = p_i = \Pr(Y=1)$ 라고 생각해 보면 종속 변수는 어떤 사건이 일어날 확률이 ($Y=1$) 된다. 여기에 odds 개념을 적용하여 종속변수를 Odds 변환을 해 보자. $p_i^* = \frac{p_i}{1-p_i}$

확률 p_i가 (0,1) 사이의 값을 가지므로 p_i^*는 (0, ∞) 값을 가진다. $\ln(p_i^*)$ 변환을 하면 이 변수는 (-∞,∞) 값을 가지므로 아래 모형에서 오차항의 $e_i \sim Normal(0,\sigma^2)$ (회귀 분석 가정)에는 문제가 없을 것이다. 이 모형을 Logistic 모형이라 한다.

$$y_i = \ln(\frac{p_i}{1-p_i}) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + .. + \beta_p x_{pi} + e_i \,,\ \ e_i \sim Normal(0,\sigma^2) \ \text{--- (로지스틱 모형)}$$

위의 모형을 다시 쓰면 다음과 같다.

$$p_i = \Pr(Y=1 \mid \underline{x}) = \frac{e^{\{\alpha+\beta_1 x_{1i}+\beta_2 x_{2i}+..+\beta_p x_{pi}\}}}{1+e^{\{\alpha+\beta_1 x_{1i}+\beta_2 x_{2i}+..+\beta_p x_{pi}\}}} + e_i^* = \frac{1}{1+e^{-\{\alpha+\beta_1 x_{1i}+\beta_2 x_{2i}+..+\beta_p x_{pi}\}}} + e_i^*$$

그러므로 회귀 계수의 부호가 양수이고 값이 커지면 p_i (성공: $Y=1$, event)가 커지므로 성공 확률이 높아지고 부호가 음수이고 절대값이 커지면 p_i가 작아지므로 성공 확률이 낮아진다.

9.1.4 모형의 적합성 검정 및 회귀계수 유의성 검정

모형 전체의 유의성은 –2Log L , AIC(Akaike Information Criterion) Schwartz Criterion을 이용하고 (Adjusted 결정계수와 유사 개념) 회귀계수의 유의성 검정은 Wald의 Chi-square 검정통계량을 이용한다.

9.2 예제 데이터 이용한 로지스틱 분석

EXAMPLE **Remission.txt** 자료는 환자의 상태를 나타내는 변수 (cell, smear, infil, li, blast, temp)들이 암 재발 여부(REMISS, 종속변수)에 영향을 미치는지 알아보기 위하여 수집한 자료이다.

9.2.1 데이터 읽기

```
data Remission;
     input remiss cell smear infil li blast temp;
     datalines;
   1   .8   .83   .66   1.9   1.1      .996
   1   .9   .36   .32   1.4    .74     .992
   0   .8   .88   .7     .8   .176    .982
```

remiss=1이면 재발

remiss=0이면 재발하지 않음.

종속변수 REMISS, 관측치는 0 혹은 1(성공)을 갖는 이진형(binary, dichotomous) 변수이다. 1이면 암 재발, 0이면 재발하지 않음.

9.2.2 OLS 추정치 문제점

OLS의 문제점을 잘 파악하기 위하여 설명변수가 하나(Li만)인 단순회귀모형을 중심으로 살펴보기로 한다. 다음은 OLS 추정방법에 의해 단순회귀모형을 추정한 결과이다.

```
proc reg data=Remission;
     model remiss=li;
     title h=1.5 'scatter plot of remisss vs. li';
     plot remiss*li;
     title h=1.5 'residual plot';
     plot residual.*predicted.;
     output out=out1 p=yhat_o;
run;
```

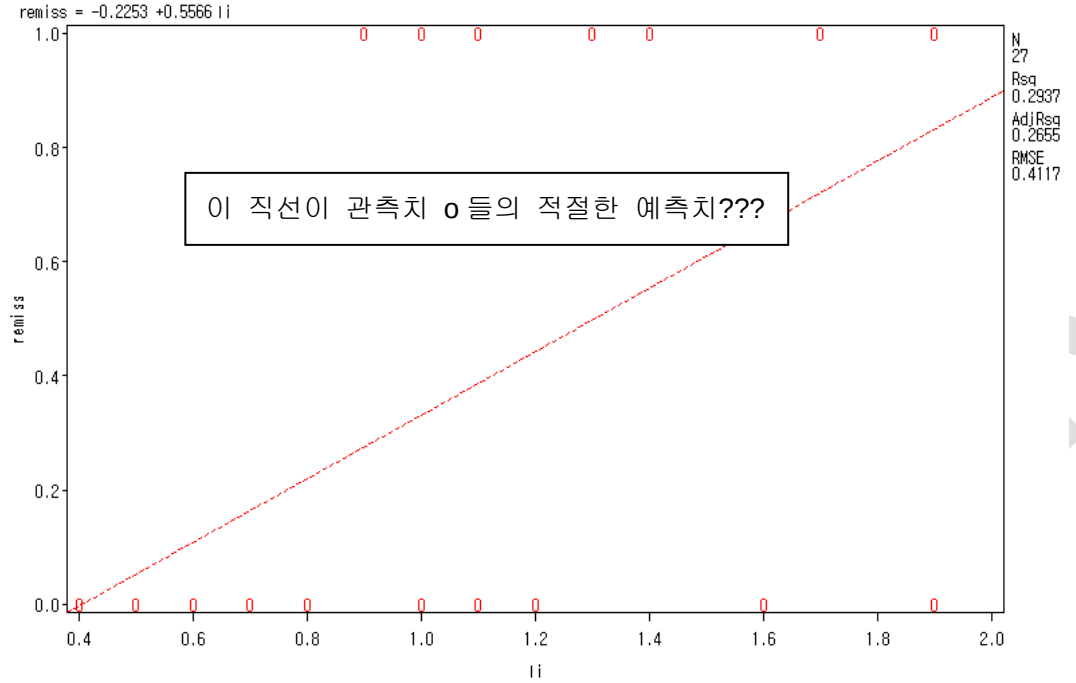
설명변수 Li는 유의하다. 그러나 종속변수가 이진형(연속형이 아니므로)이므로 결정계수의 값이 매우 낮다. 어찌되었던 결과만 본다면 우리는 $\mathrm{Re}\,miss = -0.225 + 0.557 * Li$ 모형은 유의하며 설명변수 Li는 병 재발에 양의 영향을 미친다고 결론 내릴 수 있다. 그러나 다음 페이지 원 변수 산점도와 잔차 산점도를 보면 OLS 추정 방법은 옳은 방법이 아님을 알 수 있다.

Root MSE	0.41171	R-Square	0.2937
Dependent Mean	0.33333	Adj R-Sq	0.2655
Coeff Var	123.51163		

Parameter Estimates

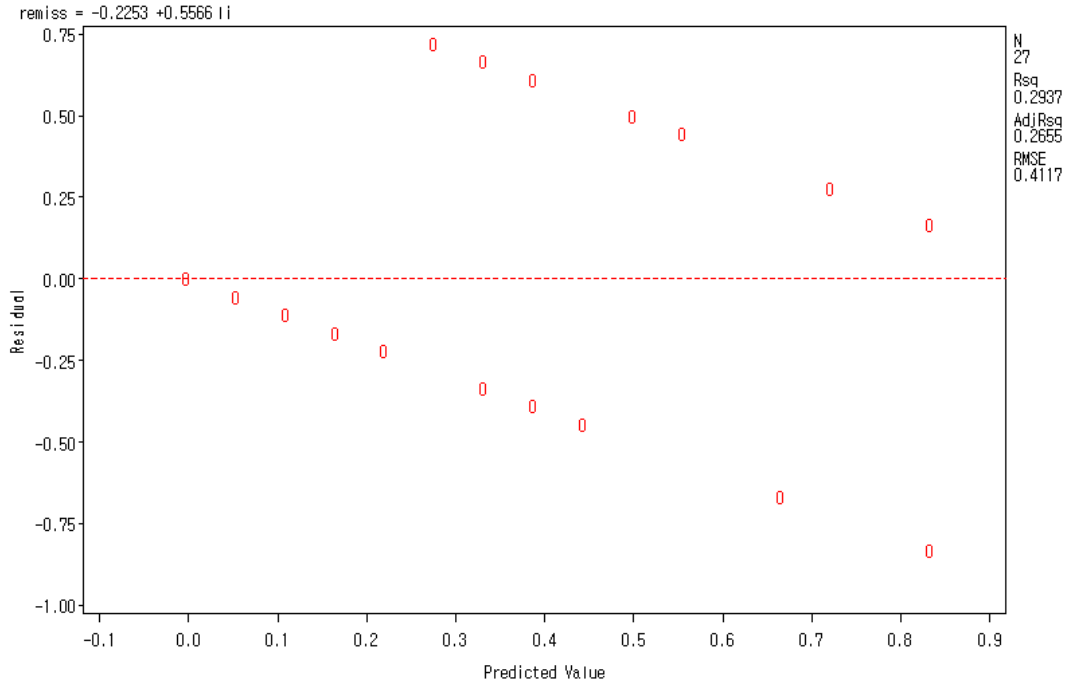
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.22530	0.19050	-1.18	0.2481
li	1	0.55657	0.17260	3.22	0.0035

scatter plot of remisss vs. li



우와 잔차들이 예술이네...

residual plot



9.2.3 로지스틱 회귀분석(설명변수 하나)

OLS 분석 방법과 비교하기 위하여 설명 변수가 Li 하나인 경우 로지스틱 분석을 실시해 보자.

```
proc logistic descending data=Remission;
     model remiss=li;
     output out=out2 p=yhat_l;
run;
```

Descending 옵션을 사용하는 이유는 SAS는 코딩 값이 작은 것을 event(사건, 성공)라 보고 큰 것을 non-event라 본다. 그런데 예제 자료는 1이 재발(이것이 event에 해당)이므로 자료 코딩을 반대로 인식하라는 명령으로 descending을 사용한다. OUTPUT 문에 의해 로지스틱 회귀 모형 추정에 의한 예측치($\hat{Y}$) 결과를 OUT2에 저장했다.

Response Profile

Ordered Value	remiss	Total Frequency
1	1	9
2	0	18

Event(성공)
non-event(실패)

Descending 옵션에 의해 REMISS 값이 큰 1이 event(성공)가 되었고 0이 non-event가 되었다. 만약 descending 옵션을 사용하지 않으면 profile 출력 결과는 다음과 같다.

Response Profile

Ordered Value	remiss	Total Frequency
1	0	18
2	1	9

Probability modeled is remiss=0.

Probability modeled is remiss=0의 의미는 event (재발)을 0(재발 않음)으로 간주했다는 의미이다.

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	8.2988	1	0.0040
Score	7.9311	1	0.0049
Wald	5.9594	1	0.0146

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-3.7771	1.3786	7.5064	0.0061
li	1	2.8973	1.1868	5.9594	0.0146

로지스틱 회귀 분석에서는 회귀 계수의 유의성 검정은 χ^2-검정에 의한다. p-값이 0.0146 이므로 Li 설명 변수는 매우 유의하다. 종속변수가 0과 1인 경우 로지스틱 회귀모형에 의한 종속변수는 $Y_i = p_i = \Pr(Y=1|\underline{x})$ (주어진 설명변수에서 사건(event, 성공) 발생 확률)로 0과 1사이의 값이다. Li 값이 커질수록 재발 가능성(확률)은 높아진다.

$$\hat{y}_i = \hat{p}_i = \Pr(Y = \text{Re}\,miss\text{재발} \mid \underline{x}) = \frac{1}{1+e^{-(3.77+2.89Li)}}$$

9.2.4 로지스틱 회귀분석(설명변수 2개 이상)

변수선택

일반 다중선택과 같이 변수선택이 가능하다. 로지스틱 분석에서 다중공선성 문제는 상관이 없다. 다음은 stepwise 변수 선택 방법에 의해 변수를 선택한 결과이다. SLS와 SLE을 다소 높게 잡은 것은 이렇게 하지 않으면 변수가 하나만 선택되기 때문이다. 이것은 예제니까 이렇게 하지 실제로는 SLS=0.15~0.2)와 SLE=0.1으로 설정하는 것이 바른 방법이다.

```
PROC LOGISTIC descending DATA=REMISSION;
    MODEL REMISS=cell smear infil li blast temp
          /selection=stepwise sls=0.3 sle=0.5;
RUN;
```

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	li		1	1	7.9311	.	0.0049
2	temp		1	2	1.2591	.	0.2618
3	cell		1	3	1.4700	.	0.2254

Analysis of Maximum Likelihood Estimates

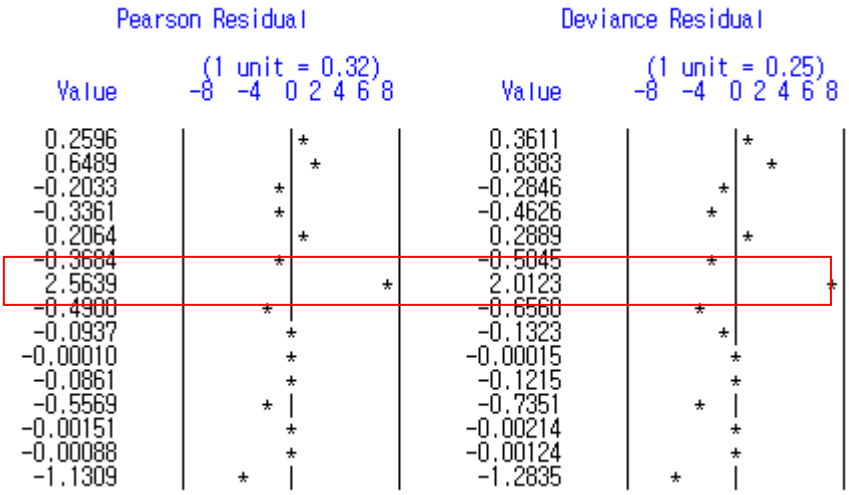
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	67.6339	56.8875	1.4135	0.2345
cell	1	9.6521	7.7511	1.5507	0.2130
li	1	3.8671	1.7783	4.7290	0.0297
temp	1	-82.0737	61.7124	1.7687	0.1835

추정회귀모형은 $\hat{y}_i = \hat{p}_i = \Pr(Y = \text{Re}\,miss\text{재발} \mid \underline{x}) = \frac{1}{1+e^{-(67.6+9.65Cell+3.87Li-82.07Temp)}}$ 로 Cell이 커질수록, Li가 커질수록, Temp가 낮아질수록 재발 가능성은 높아진다.

이제 영향치(이상치) 진단을 해 보자. 유의한 변수만을 가지고 다음 작업을 하면 된다.

```
PROC LOGISTIC descending DATA=REMISSION;
    MODEL REMISS=cell li temp /influence;
RUN;
```

영향치 관련 여러 검정통계량이 출력되나 Pearson Residuals이나 Deviance residual을 이용하면 된다. 이 값이 ±2 보다 크면 이상치로 판단하여 제거하면 된다. 제거 방법은 수작업을 한다. 수작업? 원 데이터에서 제외한다. 이상치를 제외하면 다시 이상치가 계속 발생하는 경우 제거 기준을 다소 올려 ±2.5 값을 사용한다.



로지스틱 회귀분석에서는 잔차분석(독립성, 등분산성, 정규성, 스튜던트 잔차의 산점도) 검정은 하지 않는다.

예측치 구하기 (성공 확률)

예측치(event 성공확률, 예제에서는 재발 확률)를 구하려면 다음과 같이 하면 된다. 위에서 잔차가 2.56이고 이것을 제거하기 시작하면 계속 제거되어 이상치를 제거하지 않았다.

```
PROC LOGISTIC descending DATA=REMISSION;
    MODEL REMISS=cell li temp;
    OUTPUT OUT=OUT1 P=PRED;
RUN;

PROC PRINT DATA=OUT1;
RUN;
```

remiss	cell	smear	infil	li	blast	temp	_LEVEL_	PRED
1	0.80	0.83	0.66	1.9	1.100	0.996	1	0.88955
1	0.90	0.36	0.32	1.4	0.740	0.992	1	0.72159
0	0.80	0.88	0.70	0.8	0.176	0.982	1	0.02439
0	1.00	0.87	0.87	0.7	1.053	0.986	1	0.20799

마지막의 PRED 변수는 $\hat{y}_i = \hat{p}_i = \Pr(Y = \text{Re}\,miss\text{재발} \mid \underline{x}) = \frac{1}{1+e^{-(67.6+9.65Cell+3.87Li-82.07Temp)}}$ 에 의해 예측된 확률이다. 즉 첫번째 관측치의 경우 재발 확률은 89.96%이다.

표준화 회귀계수

어느 설명변수가 더 많은 재발 확률에 영향을 주는지 알아보기 위해서는 일반 회귀분석처럼STB 옵션을 사용하면 된다.

```
PROC LOGISTIC descending DATA=REMISSION;
    MODEL REMISS=cell li temp/stb;
    OUTPUT OUT=OUT1 P=PRED;
RUN;
```

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	164.6	96.8576	2.8895	0.0892	
cell	1	19.4869	13.2853	2.1515	0.1424	2.0754
li	1	7.7137	3.4437	5.0172	0.0251	2.0023
temp	1	-193.6	108.6	3.1761	0.0747	-1.6343

CELL 변수와 LI 변수의 영향력은 유사하며 TEMP 변수의 영향력은 다소 낮다.

새로운 관측치에 예측

데이터 마지막 라인 종속변수는 .(결측치)로 하고 설명변수의 값을 넣고 위의 작업을 반복하면 예측치(예측 확률)를 구할 수 있다. (자세한 내용은 9.3절 참고)

오분류 표 출력하기

```
proc logistic descending data=remission;
    model remiss=cell li temp /stb ctable;
run;
```

Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Percentages Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.000	9	0	18	0	33.3	100.0	0.0	66.7	.
0.030	9	3	15	0	44.4	100.0	16.7	62.5	0.0
0.400	6	15	3	3	77.8	66.7	83.3	33.3	16.7
0.420	6	15	3	3	77.8	66.7	83.3	33.3	16.7
0.440	6	15	3	3	77.8	66.7	83.3	33.3	16.7
0.460	6	15	3	3	77.8	66.7	83.3	33.3	16.7
0.480	6	15	3	3	77.8	66.7	83.3	33.3	16.7
0.500	6	15	3	3	77.8	66.7	83.3	33.3	16.7
0.520	6	15	3	3	77.8	66.7	83.3	33.3	16.7
0.540	4	15	3	5	70.4	44.4	83.3	42.9	25.0
0.560	4	15	3	5	70.4	44.4	83.3	42.9	25.0

결과 해석 방법은 206페이지 다음에 200-1로 되어 있음.

9.3 종속변수가 순서형인 경우

II EXAMPLE II FITNESS.txt 자료에서 1.5마일을 달린 후 산소량(OXYGEN)은 나이, 몸무게, 달린 시간이 영향을 미칠 것이라 판단해 보자. 산소량을 순서형 변수(ordinal)로 만들기 위하여 RANK 절차를 사용해 보자.

```
PROC RANK DATA=FITNESS OUT=OUT1 GROUPS=3;
    VAR OXYGEN;
    RANKS OXYGEN_G;
RUN;

PROC PRINT DATA=OUT1;
RUN;
```

변수 OXYGEN은 크기에 의해 3개의 그룹(0, 1, 2)으로 나누어졌다.

Obs	Age	Weight	Oxygen	Run Time	Rest Pulse	Run Pulse	Max Pulse	OXYGEN_G
1	44	89.47	44.609	11.37	62	178	182	0
2	40	75.07	45.313	10.07	62	185	185	0
3	44	85.84	54.297	8.65	45	156	168	2
4	42	68.15	59.571	8.17	40	166	172	2
5	38	89.02	49.874	9.22	55	178	180	2
6	47	77.45	44.811	11.63	58	176	176	0
7	40	75.98	45.681	11.95	70	176	180	1
8	43	81.19	49.091	10.85	64	162	170	1

이제 나이, 몸무게 달린 거리가 OXYGEN_G에 영향을 미치는지 로지스틱 회귀분석을 실시해 보자. 우선 변수 선택을 해 보자.

```
PROC LOGISTIC DATA=OUT1;
    MODEL OXYGEN_G=AGE WEIGHT RUNTIME/SELECTION=STEPWISE SLS=0.2 SLE=0.1;
RUN;
```

Response Profile

Ordered Value	OXYGEN_G	Total Frequency
1	0	10
2	1	11
3	2	10

Probabilities modeled are cumulated over the lower Ordered Values.

로지스틱에서는 이 부분을 항상 체크하라.

순서 값이 가장 작은 순서대로 누적 확률이 계산된다는 의미이다.

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept 0	1	-23.7183	6.4620	13.4722	0.0002
Intercept 1	1	-20.7029	5.9822	11.9767	0.0005
RunTime	1	2.1101	0.5913	12.7358	0.0004

설명변수 중 유의한 것은 RUNTIME이었다. 이제 추정 모형을 적어보자.

$$\Pr(Oxygen\_G = 0 \mid \underline{x}) = \frac{1}{1+e^{-(-23.7+2.11Runtime)}}$$ (산소량이 가장 그룹에 속할 확률)

$$\Pr(Oxygen\_G \le 1 \mid \underline{x}) = \frac{1}{1+e^{-(-20.7+2.11Runtime)}}$$ (산소량 가장 작은&중간 그룹에 속할 누적확률)

그러므로 $\Pr(Y = 2 \mid \underline{x}) = 1 - \Pr(Oxygen\_G \le 1)$

이것에 의해 예측치를 구하면 다음과 같다. 범주가 3개 이상인 경우에는 영향치 진단이 가능하지 않다.

```
PROC LOGISTIC DATA=OUT1;
    MODEL OXYGEN_G=RUNTIME;
    OUTPUT OUT=OUT2 P=PRED;
RUN;

PROC PRINT DATA=OUT2;
RUN;
```

Run Time	Rest Pulse	Run Pulse	Max Pulse	OXYGEN_G	_LEVEL_	PRED
11.37	62	178	182	0	0	0.56789
11.37	62	178	182	0	1	0.96404
10.07	62	185	185	0	0	0.07800
10.07	62	185	185	0	1	0.63310
8.65	45	156	168	2	0	0.00421
8.65	45	156	168	2	1	0.07938
8.17	40	166	172	2	0	0.00153

종속변수 범주가 3개이므로 두 줄이 각 관측치에 대한 예측 결과이다. 첫번째 행이 0그룹에 속할 확률이고 두번째 행이 0과1에 속할 누적확률이다. 그러므로 첫번째 관측치(runtime이 11.37)이 산소량이 적은 그룹에 속할 확률은 0.57이고 중간 그룹에 속할 확률은 (0.96-0.57) =0.39, 그리고 가장 높은 그룹에 속할 확률은 1-0.96=0.04이므로 첫번째 그룹은 산소량이 가장 적은 0그룹에 속할 확률이 가장 높다.

이제 달린 시간에 11초인 경우 산소량 그룹을 예측해 보자. 다른 설명변수가 유의하지 않으므로 모두 결측치로 하고 RUNTIME만 11을 입력한다. 그리고 프로그램을 다시 실행하면 예측 확률이 구해진다.

```
   52 82.78 47.467 10.50 53 170 172
    . . . 11 . . .
   ;
run;
```

Run Time	Rest Pulse	Run Pulse	Max Pulse	OXYGEN_G	_LEVEL_	PRED
9.93	49	148	155	2	1	0.56221
10.50	53	170	172	1	1	0.81044
11.00	.	.	.	.	0	0.37578
11.00	.	.	.	.	1	0.92469

산소량 0그룹에 속할 확률이 0.37, 1그룹에는 (0.92-0.38)=0.54, 2그룹에 1-0.92=0.08이므로 달린 시간이 11초인 사람은 산소량이 1그룹(중간 정도)에 속할 가능성이 가장 높다.

HOMEWORK #11 SPSS(짝수) SAS(홀수) Due 6월 1일(수)

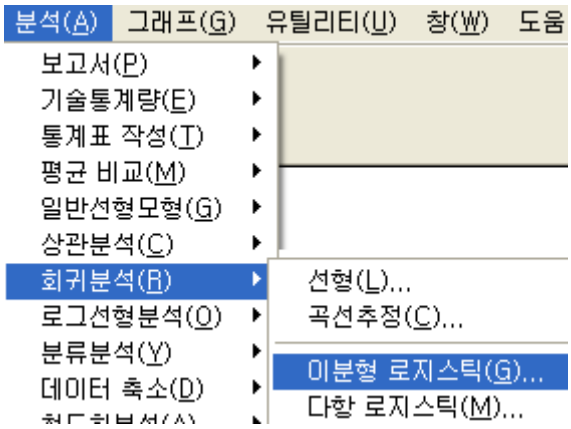
FINANCE.xls 기업의 향후 지불 능력에 재무 관련 변수가 영향을 미칠 것이라 판단했다. 그래서 X1=(보유이익/총자산), X2=(과세전 수익/총자산), X3=(매출액/총자산)을 설명변수 종속변수 Y 를 0(2 년 후 파산), 1(2 년 후 지불능력 있음)을 조사하였다.

①로지스틱 회귀분석을 실시하시오. 변수 선택, 이상치 제거, 유의한 설명변수가 2 개 이상이 되도록 유의수준을 다소 조정하시오.

②유의한 설명변수의 영향력을 비교하시오.

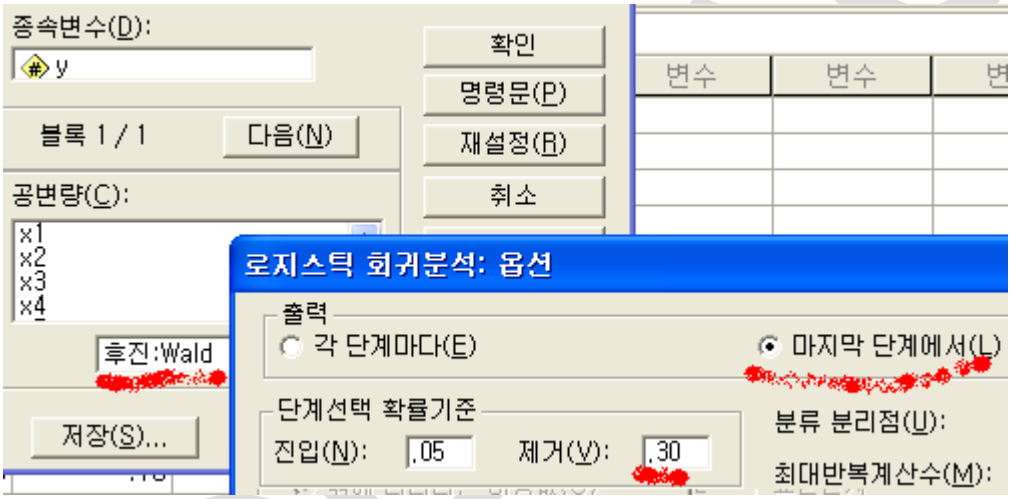
③X1=20, X2=10, X3=1 인 기업이 2 년 후 파산할 가능성(확률)을 예측하시오.

SPSS 에서 로지스틱 회귀분석하기(이진형)



종속변수 코딩

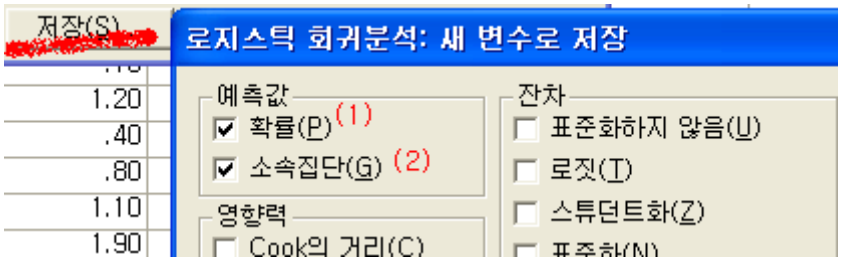
원래 값	내부 값
0	0
1	1 ←성공



방정식에 포함된 변수

	B	S.E.	Wald	자유도	유의확률	Exp(B)
X4	2.897	1.187	5.959	1	.015	18.124
상수	-3.777	1.379	7.506	1	.006	.023

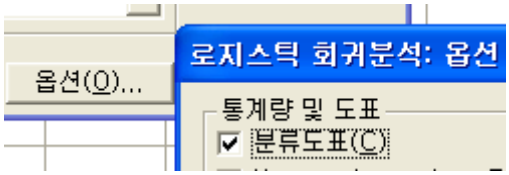
로지스틱 모형에서 종속변수는 성공이 발생할 (1)확률 $\hat{y}_i = \Pr(Y=1|\underline{x})$ 이므로 0과 1 사이의 값이다. 로지스틱 회귀분석 창의 "저장" 옵션을 선택하여 예측 값을 아래와 같이 설정하면 된다. (2)소속집단을 선택하면 집단 분류를 나타내는 변수가 생성된다. 집단(성공/실패) 분류 기준은 성공 확률 0.5이다.



PRE_*이 성공 확률 예측값이고 PGR_*는 그룹 정보 변수이다. X4=1.9인 환자는 병이 재발할 확률은 0.849이다. 그러므로 cut-off을 0.5라 하면 병이 재발할 환자 그룹으로 분류할 수 있다.

y	x4	x5	x6	pre_1	pgr_1
1	1.90	1.10	1.00	.84911	1
1	1.40	.74	.99	.56931	1
0	.80	.18	.98	.18857	0
0	.70	1.05	.99	.14817	0

분류 기준을 0.5로 하여 분류한 결과와 원 그룹 결과에 대한 표는 다음과 같다. 로지스틱 회귀모형에 의해 개체를 분류하면 7개의 오분류가 생긴다. 특히 재발 환자에 대한 예측이 44%로 매우 낮으므로 좋은 모형은 아니다. 로지스틱 분석을 개체 판별에 이용하려면 오분류 표에 의해 cut-off는 분석자가 임의로 조정할 수 있다.



분류표

관측			예측값 Y 0	예측값 Y 1	분류정확 %
1 단계	Y	0	16	2	88.9
		1	5	4	44.4
	전체 %				74.1

일반적으로 로지스틱 분석에서는 산점도(잔차와 예측치 등), 잔차의 정규성을 이용한 잔차분석은 실시하지 않는다. 로지스틱 회귀모형에서 지시변수 사용 방법은 일반 회귀분석과 동일하게 분류형 설명변수의 수준에 따라 지시변수(들)을 만들어 사용하며 된다.

새로운 개체에 대한 성공 확률 예측은 일반 선형 회귀모형과 동일하다. X4가 1.25인 환자는 병의 재발 확률이 0.46이다.

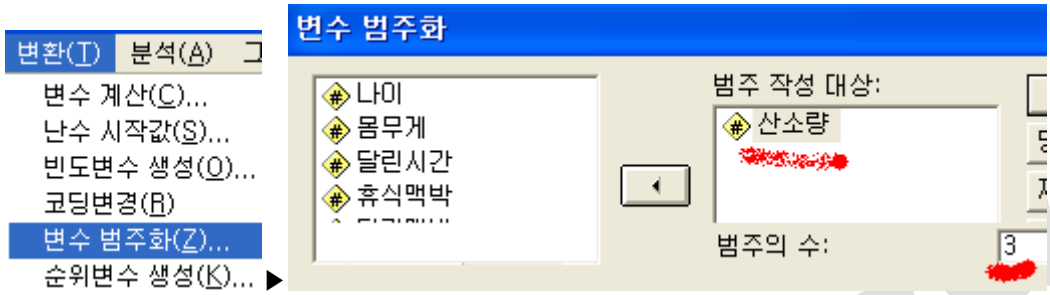
y	x1	x2	x3	x4	pre_2	pgr_2
.	.	.	.	1.25	.46119	0

SPSS 에서 로지스틱 회귀분석하기(순서형)

변수명: 나이, 몸무게, **산소량**, 달린 시간, 휴식 맥박, 달릴 때 맥박, 최대 맥박 (입력 순)

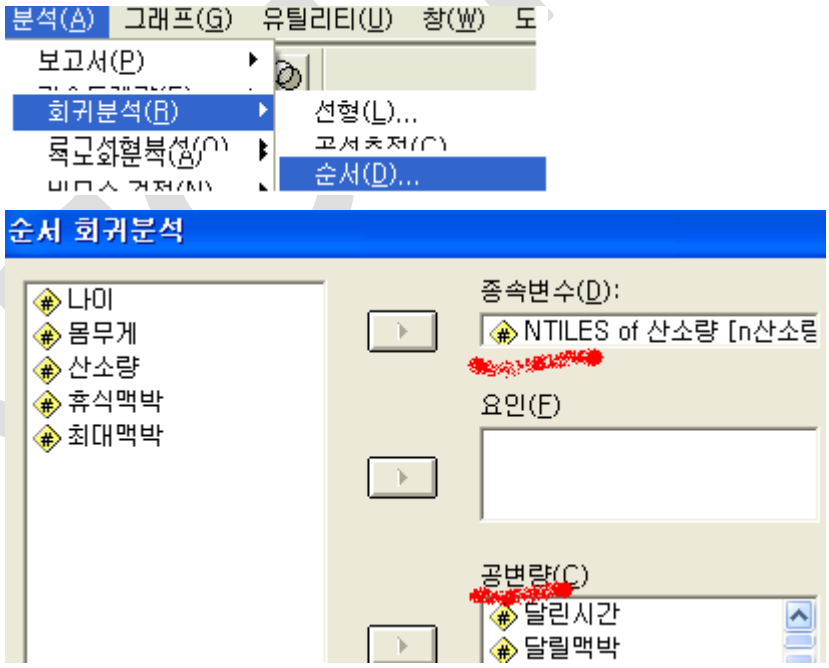
▶산소량이 측정형 변수이나 이를 순서형 변수로 만들어 종속변수로 사용하자.

우선 측정형 변수를 순서형(범주형) 변수로 만들자. 아래 방법에 의해 3개의 그룹(산소량 상, 중, 하)으로 만들었다.



산소량	달린시간	휴식맥박	달릴맥박	최대맥박	n산소량
44.61	11.37	62.0	178.00	182.00	1
45.31	10.07	62.0	185.00	185.00	1
54.30	8.65	45.0	156.00	168.00	3
59.57	8.17	40.0	166.00	172.00	3
49.87	9.22	55.0	178.00	180.00	3
44.81	11.63	58.0	176.00	176.00	1
45.68	11.95	70.0	176.00	180.00	2
49.09	10.85	64.0	162.00	170.00	2

순서형 로지스틱 회귀분석 절차이다. 종속변수와 설명변수(공변량)들을 선택하고 다른 옵션들은 default 옵션으로 하여 회귀모형을 추정한다. 순서형 로지스틱 분석에는 변수 선택 옵션이 없으므로 가장 유의하지 않은 설명변수(몸무게)부터 하나씩 수작업으로 제거한다.



모수 추정값

		B 추정값	표준 오차	Wald	PAR 유의확률	95% 신뢰구간 하한	95% 신뢰구간 상한
한계치	[N산소량 = 1]	-49,069	20,553	5,700	,017	-89,352	-8,786
	[N산소량 = 2]	-45,469	20,062	5,137	,023	-84,789	-6,148
위치	달린시간	-2,607	,856	9,270	,002	-4,286	-,929
	최대맥박	,160	,199	,647	,421	-,230	,551
	휴식맥박	4,904E-02	,077	,407	,524	-,102	,200
	달릴맥박	-,264	,185	2,039	,153	-,626	9,824E-02
	몸무게	-2,62E-02	,063	,170	,680	-,150	9,806E-02
	나이	-7,25E-02	,118	,379	,538	-,303	,158

링크 함수: 로짓.

몸무게▶나이▶휴식 맥박▶최대 맥박 순으로 제거하면 다음 추정 결과를 얻는다.

모수 추정값

		B 추정값	표준 오차	Wald	PAR 유의확률	95% 신뢰 상한
한계치	[N산소량 = 1]	-43,211	13,230	10,669	,001	-17,282
	[N산소량 = 2]	-39,688	12,633	9,870	,002	-14,928
위치	달린시간	-2,308	,679	11,555	,001	
	달릴맥박	-,101	,052	3,792	,052	

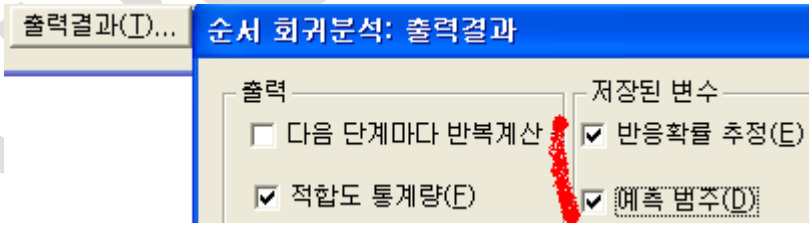
링크 함수: 로짓.

$$\Pr(Oxygen\_G = 1 \mid \underline{x}) = \frac{1}{1+e^{-(-43.2-2.31Runtime-1.01RunPulse)}}$$ (산소량이 가장 작은 그룹)

$$\Pr(Oxygen\_G \le 2 \mid \underline{x}) = \frac{1}{1+e^{-(-39.7-2.31Runtime-0.101RunPulse)}}$$ (가장 작은 그룹, 중간 그룹)

그러므로 $\Pr(Y = 3 \mid \underline{x}) = 1 - \Pr(Oxygen\_G \le 2)$ (산소량 가장 높은 그룹)

달린 시간이 오랠수록, 달릴 때 맥박이 높을수록 산소량은 줄어들고 있음을 알 수 있다. 각 개체에 대한 그룹에 속할 확률(변수명은 EST*)은 다음 방법에 의해 추정된다. 예측 범주는 확률이 가장 높은 집단으로 할당된다.



n산소량	est1_1	est2_1	est3_1	pre_1
1	.74	.25	.01	1
1	.22	.68	.09	2
3	.00	.02	.98	3
3	.00	.02	.98	3
3	.02	.39	.60	3
1	.81	.18	.01	1

Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Correct	Percentages Sensi-tivity	Speci-ficity	False POS	False NEG
0.000	9	0	18	0	33.3	100.0	0.0	66.7	.
0.030	0	,3	11	0	59.3	100.0	38.0	55.0	,0.0
0.400	6	15	3	3	77.8	66.7	83.3	33.3	16.7
0.420	6	15	3	3	77.8	66.7	83.3	33.3	16.7
0.440	6	15	3	3	77.8	66.7	83.3	33.3	16.7
0.460	6	15	3	3	77.8	66.7	83.3	33.3	16.7
0.480	6	15	3	3	77.8	66.7	83.3	33.3	16.7
0.500	6	15	3	3	77.8	66.7	83.3	33.3	16.7
0.520	6	15	3	3	77.8	66.7	83.3	33.3	16.7
0.540	4	15	3	5	70.4	44.4	83.3	42.9	25.0
0.560	4	15	3	5	70.4	44.4	83.3	42.9	25.0

①Prob. Level은 Logit 회귀분석에 의해 추정된 Pr(Event) 확률이 이 값 이상이면 Event로 정의한다는 의미이다. 첫 열을 보면 추정된 Pr(Event) 확률이 0.0 이상이면 Event로 분류한다면 가정하고 개체를 분류한 결과를 보여준다. 그러므로 EVENT 19개는 모두 Event로 정분류 되고 non-EVENT 18개는 모두 Event로 분류되므로 오분류 된다.

②Correct Event(Domestic) 칠면조를 Event로 Non-event 칠면조를 Non-event로 정분류

③In-Correct Event(Domestic) 칠면조를 non-Event로 Non-event 칠면조를 event로 오분류

④Correct는 전체 개체 수 중 정분류 된 개체 비율을 출력한 것이다.

⑤Sensitivity는 Event(사육) 개체를 Event(사육)으로 정분류 비율

⑥False Pos. 는 Event(사육) 개체를 non-Event(야생)으로 오분류 비율, 그러므로 Sensitivity + False Pos. 는 1이다.

⑦Specificity는 non-Event(야생) 개체를 non-Event(야생)으로 정분류 비율

⑧False Neg. 는 non-Event(야생) 개체를 Event(사육)으로 오분류 비율, 그러므로 Specificity + False Neg. 는 1이다.

CORECT가 가장 적은 Prob. Level 기준 값을 이용하여 개체를 분류하면 된다.