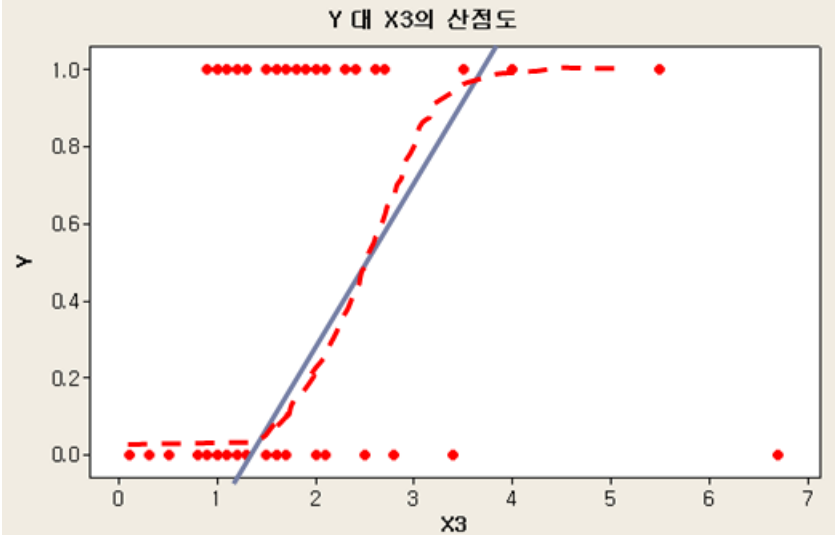
- 종속변수가 이진(binary: 가질 수 있는 값이 실패/성공, 정품/불량 등과 같이 가질 수 있는 값이 2 개인 경우)인 경우 => 로지스틱 회귀분석(Logistic regression)
- 로지스틱 회귀분석에서 종속변수 값은 0, 1(사건: 성공, 불량)로 입력된다.
- 로지스틱 회귀분석은 이진형 반응변수 뿐 아니라 반응변수가 순서형(ordinal) 분류형인 경우 사용할 수 있다. 예를 들면 종속 변수가 고객의 신용도이고 이 변수가 (상, 중, 하) 분류되어 있는 경우 사용할 수 있다.
- 종속변수의 수준이 3 개 이상인 경우 LOGISTIC 모형을 사용하는 것이 아니라 CATMOD 를 사용해야 한다고 언급한 책이 있다. 그러나 CATMD 는 CATegorical data MODeling 의 약어로 분류변수 자료 모형화이며, LOGISTIC 모형은 CATMOD 기법의 한 부분이다.

로지스틱 모형

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 0 과 1 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, 1 만 존재) OLS 에 의한 계수 추정은 무의미 하여 음의 값이 예측되거나 부호 자체가 달라지게 된다. 또한 이진형 변수의 특성상 이분산 가능성이 매우 높다.



o 회귀계수 부호와 로지스틱 회귀계수는 부호는 일치한다.

o 그리고 유의성도 어느 정도 일치함

Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

ODDS

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

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_i 가 (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.

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

위의 모형을 다시 쓰면 다음과 같다.

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

$$= \frac{1}{1+e^{-\{\alpha+\beta_1 x_{1i}+\beta_2 x_{2i}+..+\beta_p x_{pi}\}}} + e_i^*$$

o 회귀 계수의 부호가 양수이고 값이 커지면 p_i (성공: $Y=1$, event)가 커지므로 성공 확률이 높아지고 부호가 음수이고 절대값이 커지면 p_i 가 작아지므로 성공 확률이 낮아진다.
o Exp(b)는 오즈비에 미치는 설명변수의 한 단위 곱의 영향

모형의 적합성 검정 및 회귀계수 유의성 검정

o 모형 전체의 유의성은 –2Log L , AIC(Akaike Information Criterion) Schwartz Criterion (Adjusted 결정계수와 유사 개념)
o 회귀계수의 유의성 검정은 Wald 의 Chi-square 검정통계량 (⇔ z-통계량)

Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

예제 데이터

🖫 Remission.csv 자료는 환자의 상태를 나타내는 변수 (cell, smear, infill, lithium, blast, temp)들이 암 재발 여부(REMISS, 종속변수 No=1, Yes=2)에 영향을 미치는지 알아보기 위하여 수집한 자료이다.

분석

```
> fit.re=glm(remiss~cell+smear+infill+lithium+blast+temp,data=ds.re,
+ family=binomial(link = "logit"))
> summary(fit.re) # display results
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  24.5814    66.2720   0.371    0.711
cell         59.8173    54.2902   1.102    0.271
smear        70.4212    65.8708   1.069    0.285
infill      -72.3608    69.3947  -1.043    0.297
lithium       3.4552     2.2400   1.542    0.123
blast         0.2317     2.2071   0.105    0.916
temp        -87.5152    64.3248  -1.361    0.174
                                          AIC: 36.728
```

o 유의하지 않은 설명변수를 회귀분석과 동일하게 수작업으로 하나씩 제외, backward

```
> fit.re=glm(remiss~lithium+temp,data=ds.re,family=binomial(link = "logit"))
> summary(fit.re) # display results

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    62.72      46.21   1.357    0.175
lithium         3.07       1.33   2.308    0.021 *
temp          -66.90      47.28  -1.415    0.157
                                          AIC: 32.148
```

o 설명변수 Lithium 부호가 양이므로 환자의 재발 (yes 가 no 에 비해 크므로 재발이 “성공” 개념) 확률을 높임

o 설명변수 Temp 부호가 음이므로 재발 가능성을 낮춤

```
> exp(coef(fit.re)) # exponentiated coefficients
  (Intercept)      lithium         temp
1.737973e+27 2.154564e+01 8.819609e-30
```

o 설명변수 Lithium 의 EXP(b)가 1 보다 크므로 Lithium 값이 높은 환자는 재발하지 않을 확률 대비 재발 확률이 높음.

o 설명변수 Temp 의 EXP(b)가 1 보다 작으므로 Lithium 값이 낮은 환자는 재발하지 않을 확률 대비 재발 확률이 높음.

Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

예측확률

```
> predict(fit.re, type="response") # predicted values
          1           2           3           4           5
0.872458540 0.658223886 0.373388089 0.251180023 0.759726164
```

추정된 로지스틱 회귀모형 $p_i = \Pr(Y=1|\underline{x}) = = \frac{1}{1+e^{-\{62.7+3.07*Li-66.9*Temp\}}}$ 에 의해 추정된 환자의 재발확률이다.

```
> p.g=predict(fit.re, type="response")>0.5 # 예측집단
> table(ds.re$remiss,p.g)
     p.g
      FALSE TRUE
  No     14    3
  Yes     4    6
```

재발 확률이 0.5 이상인 경우 환자를 재발로 분류하고자 함 (기준 설정)

정분류 확률 : 성공(실패) 환자를 추정된 로지스틱 회귀분석에 의해 성공(실패)으로 분류
⇨ 20/27
오분류 확률 : 성공(실패) 환자를 추정된 로지스틱 회귀분석에 의해 실패(성공)으로 분류
⇨ 7/27

```
> prop.table(table(ds.re$remiss,p.g),1)
     p.g
         FALSE      TRUE
  No  0.8235294 0.1764706
  Yes 0.4000000 0.6000000
```

Sensitivity (민감도) : 재발 환자를 재발로 정분류 할 확률 60%
False Pos. : 재발하지 않을 환자를 재발로 오분류 할 확률 40%

Specificity (특이도) : 재발하지 않은 환자를 재발하지 않음으로 정분류 할 확률 82.4%
False Neg. : 재발한 환자를 재발하지 않은 오분류 할 확률 17.6%

프로젝트 데이터

종속변수를 적절히 이분류하여 로지스틱분석을 실시하여 2 페이지 리포트를 작성하시오.
(오분류 개체에 대한 설명 포함)

Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr