데이터마이닝(DataMining)

Chapter 3.1. 로지스틱 회귀모형

- 로지스틱 회귀(logistic regression)모형은 반응변수가 범주형인 경우에 적용되는 회귀모형
- 새로운 설명변수(또는 예측변수)의 값이 주어질 때 반응변주의 각 범주(또는 집단)에 속할 확률이 얼마인지를 추정해 주며(예측모형), 추정 확률의 기준치에 따라 반응변수를 분류(또는 판별)하는 목적으로 사용
- 모형의 적합을 통해 추정된 확률을 사후확률(posterior probability)

로지스틱 회귀

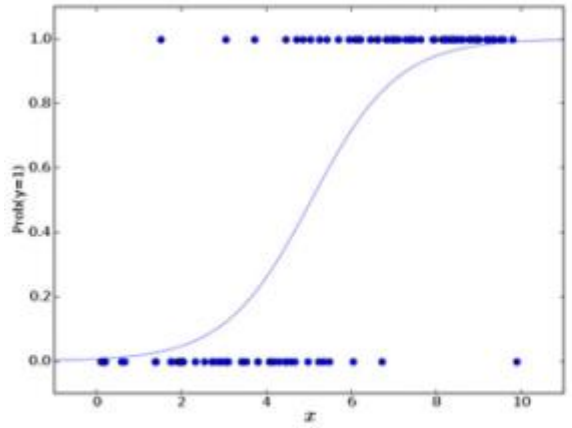
- 이항 반응변수 Y에 대해, 다중(multiple) 로지스틱 회귀모형의 일반적 형태

$$\log\left(\frac{\pi(x)}{1-\pi(x\,)}\right) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

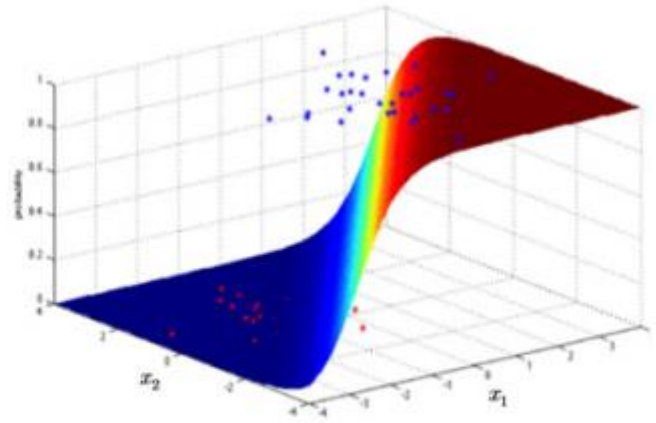
- 위 식에서 $\pi(x) = P(Y = 1|x),\ \ x = (x_1, x_2, \cdots, x_k)$
- 로지스틱 회귀모형은 오즈(odds)의 관점에서 해석될 수 있다는 장점
- $\exp(\beta_1)$의 의미는 나머지 변수 $(x_2, \cdots, x_k)$가 주어질 때 x_1이 한 단위 증가할 때마다 성공 $(Y = 1)$의 오즈가 몇 배 증가하는지를 나타내는 값
- 로지스틱 모형은 $\pi(x)$에 관한 식으로 재표현

$$\pi(x) = \frac{\exp(\alpha + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + \exp(\alpha + \beta_1 x_1 + \cdots + \beta_k x_k)} = \frac{1}{1 + \exp\{-(\alpha + \beta_1 x_1 + \cdots + \beta_k x_k)\}}$$

- 예측변수가 1개인 경우와 2개인 경우에 대해 로지스틱 회귀를 적합한 결과



(a) 예측변수가 1개인 경우



(b) 예측변수가 2개인 경우

- 식의 형태는 소위 다중로지스틱함수에 해당되며, 그래프의 형태는 설명변수가 한 개(x_1)인 경우 해당 회귀계수 β_1의 부호에 따라 S자 모양($\beta_1 > 0$) 또는 역 S자 모양 ($\beta_1 < 0$)
- 표준로지스틱 분포의 누적분포함수(c.d.f.)를 $F(x)$라 할 때

$$\pi(x) = F(\alpha + \beta_1 x_1 + \cdots + \beta_k x_k)$$

과 동일한 표현이며, 그 의미는 표준로지스틱 분포의 누적분포함수로 성공의 확률을 설명(또는 추정)한다는 의미

- 로지스틱 회귀모형과 유사한 프로빗(probit) 모형은 위 식에서 $F(\cdot)$ 대신 표준정규분포의 누적 함수 $\Phi(\cdot)$ 로 성공의 확률을 모형화한 것. 즉, 프로빗 모형은 다음과 같이

$$\Phi^{-1}\big(\pi(x)\big) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

  으로 표현

- 로지스틱 회귀가 분류의 목적으로 사용될 경우에는 $\pi(x)$가 기준값(예를 들어, 1/2)보다 크면 $Y = 1$인 집단으로, 작으면 $Y = 2$ 인 집단으로 분류
- 분류 기준값의 결정은 사전정보 또는 손실함수를 이용하거나 (정분류율, 민감도, 특이도)를 동시 에 고려하는 등의 다양한 방법이 사용

```
> data(iris)
> a <- subset(iris, Species == "setosa" | Species == "versicolor")
> a$Species <- factor(a$Species )
> str(a)
 'data.frame‘ :       100 obs. of 5 variables:
 $ Sepal.Length      :  num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width       :  num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 ...
 $ Petal.Length      :  num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 ...
 $ Petal.Width       :  num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 ...
 $ Species     :  Factor w/ 2 levels "setosa","versicolor": 1 1 1 1
1 1 1 1 1 1 ...
```

- 위 결과에서 Species는 Factor형 변수로 setosa는 $Y = 1$, versicolor는 $Y = 2$ 로 인식하고 있음 을 나타냄
- 이 자료에 대해 로지스틱 회귀가 적용될 때, 보다 큰 숫자인 versicolor일 오즈를 모형화 하므로 해석에 유의

- glm() 함수를 이용하여 로지스틱 회귀모형을 적합

```
> b <- glm(Species~Sepal.Length, data=a, family=binomial)
```

- summary() 함수를 통해 그 결과를 확인

```
> summary(b)
 Call:
 glm(formula = Species ~ Sepal.Length, family = binomial, data = a)
```

```
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-2.05501 -0.47395 -0.02829 0.39788 2.32915

 Coefficients:
              Estimate Std. Error z value Pr(>|z|)
 (Intercept)   -27.831      5.434  -5.122 3.02e-07 ***
 Sepal.Length    5.140      1.007   5.107 3.28e-07 ***
 ---
 Signif. codes:   0 ‘***’   0.001 ‘**’   0.01 ‘*’   0.05 ‘.’   0.1
‘ ’   1

 (Dispersion parameter for binomial family taken to be 1)

     Null deviance: 138.629 on 99 degrees of freedom
 Residual deviance: 64.211 on 98 degrees of freedom
 AIC: 68.211

 Number of Fisher Scoring iterations: 6
```

- 회귀계수의 검정에서 p-값이 거의 영(0)이므로 Sepal.Length가 매우 유의한 변수이며, Sepal.Length가 한 단위 증가함에 따라 versicolor(Y=2)일 오즈가 exp(5.140)≈170배 증가함

- 위 결과의 마지막 부분에 제시된 Null deviance는 절편만 포함하는 모형(H_0: $\beta = 0$하의 모형)의 완전모형(또는 포화모형)으로부터의 이탈도(deviance)를 나타내며 p-값=($\chi^2(99) > 138.629) \approx 0.005$으로 통계적으로 유의하므로 적합결여를 나타냄

- Residual deviance는 예측변수 Sepal.Length가 추가된 적합 모형의 이탈도를 나타냄. Null deviance에 비해 자유도 1 기준에 이탈도의 감소가 74.4 정도의 큰 감소를 보이며, p-값= ($\chi^2(98) > 64.211) \approx 0.997$ 이므로 귀무가설(적합 모형)이 기각되지 않으며 적합값이 관측된 자료를 잘 적합하고 있다고 말할 수 있음

```
> coef(b)
(Intercept) Sepal.Length
 -27.831451     5.140336
```

```
> exp(coef(b)["Sepal.Length"])
 Sepal.Length
     170.7732
```

- 회귀계수 β와 오즈의 증가량 $\exp(\beta)$에 대한 신뢰구간

```
> confint(b, parm = "Sepal.Length")
 Waiting for profiling to be done...
    2.5 %   97.5 %
 3.421613 7.415508
```

```
> exp(confint(b, parm = "Sepal.Length"))
 Waiting for profiling to be done...
   2.5 %     97.5 %
30.61878 1661.55385
```

- fitted() 함수를 통해 적합 결과를 확인

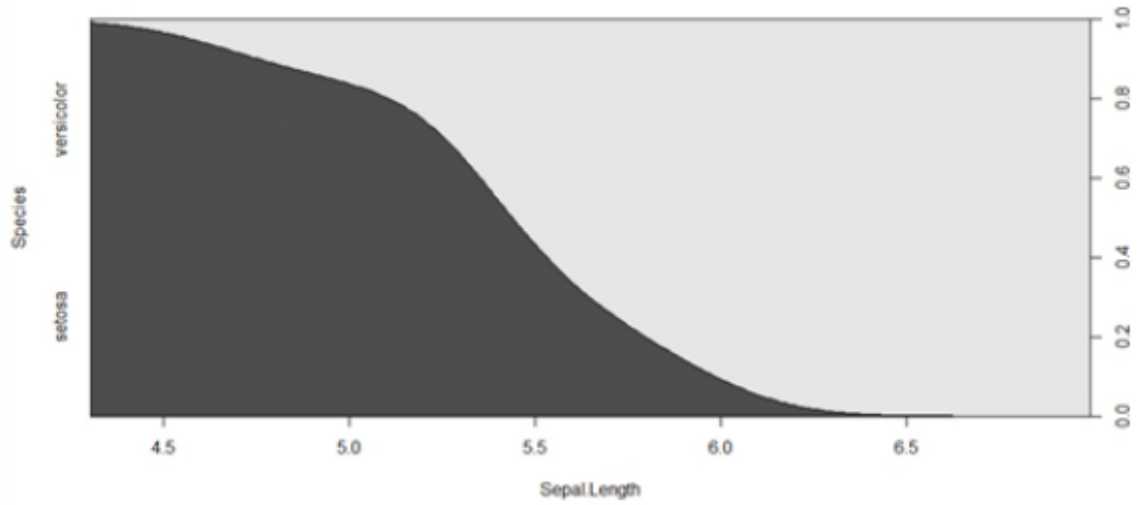
```
> fitted(b)[c(1:5, 96:100)]
         1          2          3          4          5
0.16579367 0.06637193 0.02479825 0.01498061 0.10623680
        96         97         98         99        100
0.81282396 0.81282396 0.98268360 0.16579367 0.81282396
```

- predict() 함수를 이용하여 새로운 자료에 대한 예측을 수행

```
> predict(b, newdata=a [c(1, 50, 51, 100), ], type="response" )
        1        50        51       100
0.1657937 0.1062368 0.9997116 0.8128240
```

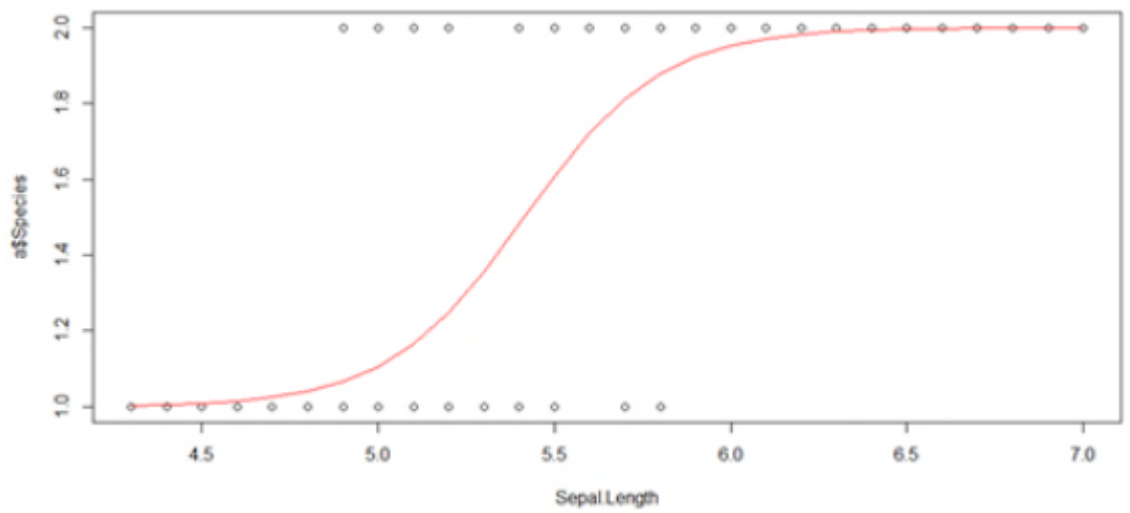
- cdplot() 함수는 Sepal.Length(연속형 변수)의 변화에 따른 범주형 변수의 조건부 분포를 보여줌
- Sepal.Length가 커짐에 따라 versicolor의 확률이 증가

```
> cdplot(Species~Sepal.Length, data=a)
```



- 적합된 로지스틱 회귀모형의 그래프

```
> plot(a$Sepal.Length, a$Species, xlab="Sepal.Length")
> x=seq(min(a$Sepal.Length), max(a$Sepal.Length), 0.1)
> lines(x, 1+(1/(1+(1/exp(-27.831+5.140*x)))), type="l", col="red")
> # 1+: 추정된 확률의 범위를 (0,1)에서 (1,2)로 조정
```



- 예측변수가 여러 개인 다중 로지스틱 회귀모형
- 1973~1974년도에 생산된 32 종류의 자동차에 대해 11개의 변수를 측정한 자료

```
> attach(mtcars)
> str(mtcars)
 'data.frame':   32 obs. of 11 variables:
 $ mpg  : num 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl  : num 6 6 4 6 8 6 8 4 4 6 ...
 $ disp : num 160 160 108 258 360 ...
 $ hp   : num 110 110 93 110 175 105 245 62 95 123 ...
 $ drat : num 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt   : num 2.62 2.88 2.32 3.21 3.44 ...
 $ qsec : num 16.5 17 18.6 19.4 17 ...
 $ vs   : num 0 0 1 1 0 1 0 1 1 1 ...    # 엔진 유형: V 엔진(0)과 straight 엔진(1)
 $ am   : num 1 1 1 0 0 0 0 0 0 0 ...
 $ gear : num 4 4 4 3 3 3 3 4 4 4 ...
 $ carb : num 4 4 1 1 2 1 4 2 2 4 ...
```

- 이항 변수 vs(0:flat engine, 1:straight engine)를 반응변수로, mpg(miles/gallon)와 am(Transmission: 0=automatic, 1: manual)을 예측변수로 하는 로지스틱 회귀모형을 적합

```
> glm.vs <- glm(vs~mpg+am, data=mtcars, family=binomial)
> summary(glm.vs)
 Call:
 glm(formula = vs ~ mpg + am, family = binomial, data = mtcars)

Deviance Residuals:
     Min       1Q   Median       3      Max
-2.05888 -0.44544 -0.08765 0.33335 1.68405
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.7051     4.6252  -2.747 0.00602 **
mpg           0.6809     0.2524   2.698 0.00697 **
am           -3.0073     1.5995  -1.880 0.06009 .
---
Signif. codes:   0 ‘***’   0.001 ‘**’   0.01 ‘*’   0.05 ‘.’   0.1
‘ ’   1

 (Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43.860 on 31 degrees of freedom
 Residual deviance: 20.646 on 29 degrees of freedom
 AIC: 26.646

 Number of Fisher Scoring iterations: 6
```

- 다중로지스틱에서 추정된 회귀계수 $\widehat{\beta_1}$에 대한 해석 : 다른 모든 변수들(여기서는 am)이 주어질 때, mpg 값이 한 단위 증가함에 따라 vs가 1일 오즈가 exp(0.6809)≈1.98배(즉, 98%) 증가
- 마찬가지로, mpg가 주어질 때, 오즈에 대한 am의 효과는 exp(-3.0073) 0.05배 즉, 변속기가 수동인 경우 자동에 비해 vs=1의 오즈가 95%나 감소

- 예측변수가 여러 개인 모형의 적합 시 변수선택법을 적용하기 위해서는 direction= 옵션을 사용한다. direction= 옵션에는 “both", "backward", "forward"가 있으며, 디폴트는 "backward"가 적용

```
> step.vs <- step(glm.vs, direction="backward")
 Start: AIC=26.65
 vs ~ mpg + am

       Df Deviance    AIC
<none>      20.646 26.646
- am    1   25.533 29.533
- mpg   1   42.953 46.953
```

- 아래의 anova() 함수는 모형의 적합(변수가 추가되는) 단계별로 이탈도의 감소량과 유의성 검정 결과를 제시

```
> anova(glm.vs, test="Chisq")
 Analysis of Deviance Table
 Model: binomial, link: logit
 Response: vs
 Terms added sequentially (first to last)

     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                    31     43.860
mpg   1   18.327        30     25.533 1.861e-05 ***
 am   1    4.887        29     20.646 0.02706 *
 ---
 Signif. codes:   0 ‘***’   0.001 ‘**’   0.01 ‘*’   0.05 ‘.’   0.1
‘ ’   1
```

- 절편항만 포함하는 영(null) 모형에서 mpg와 am 변수가 차례로 모형에 추가됨에 따라 발생하는 이탈도의 감소량을 제시하며, p-값은 각각 $P(\chi^2(1) > 18.327)$과 $P(\chi^2(1) > 4.887)$을 계산한 값
- 두 변수가 차례로 추가되면서 생겨나는 이탈도의 감소량이 모두 통계적으로 유의함을 나타냄

```
> 1-pchisq(18.327, 1)
[1] 1.860515e-05
```

```
> 1-pchisq(4.887, 1)
[1] 0.02705967
```

- 로지스틱 회귀모형은 일반화선형모형(generalized linear model)의 특별한 경우로 로짓(logit) 모형으로 부르기도 함

- 반응변수의 범주가 3개 이상인 경우에는, 범주의 유형(명목형 또는 순서형)에 따라, 다양한 다범주(multi-category) 로짓모형을 적합할 수 있음

- 일반화선형모형(GLM)에서 이탈도(deviance)
  - Null Deviance와 Residual Deviance는 각각 절편모형과 제안모형의 완전모형으로부터의 이탈도를 나타내며 다음과 같이 정의

$$LL = \text{로그가능도}$$
$$Null\ Deviance = 2\{LL(\text{포화모형}) - LL(\text{영모형})\},\ df = n - 1$$
$$Residual\ Deviance = 2\{LL(\text{포화모형}) - LL(\text{제안모형})\},\ df = n - (p + 1)$$

  - 포화모형(Saturated Model)은 추정해야 할 모수의 수가 데이터의 수와 동일한 모형으로 완전모형 (Full Model)이라고도 함
  - 영모형(Null Model)은 절편항만 가지는 모형으로 추정할 모수가 1개임
  - 제안모형(Proposed Model)은 (p개의 모수+절편항)을 포함하는 모형으로, 추정할 모수가 (p+1)개임

- Null Deviance와 Residual Deviance는 값이 작을수록 해당모형(영 또는 제안모형)이 자료를 잘 적합함을 의미

- 이탈도에 기초한 구체적인 검정은 두 종류의 Deviance가, 해당모형이 참일 때, 근사적으로 카이 제곱분포를 따른다는 사실에 기초. 이때, 자유도는 "(포화모형의 모수의 수)-(해당모형의 모 수의 수)"

- 또한, 영모형과 제안모형 간의 비교(검정)는 (Null Deviance - Proposed Deviance)가 근사적 으로 자유도가 (n-(p+1))-(n-1)=p인 카이제곱분포를 따른다는 사실에 기초