

데이터과학을 위한 **R**프로그래밍

13주차. 연관규칙과 로지스틱모형



이혜선 교수

포항공과대학교 산업경영공학과



목차

13주차. 연관규칙과 로지스틱모형

1차시 연관규칙 I

2차시 연관규칙 II

3차시 로지스틱 회귀모형

An isometric illustration of a business meeting. In the center, a large white hexagonal table is surrounded by several people. To the left, a large screen displays various charts and graphs. To the right, another screen shows a grid of data and gears. In the background, there are more people interacting with screens and a stack of red blocks. The overall scene is set on a light blue floor with a dark blue border.

13주차

3차시

로지스틱 회귀모형


로지스틱 회귀모형

Y : Remiss (0, 1)

6 explanatory variables : risk factor related cancer remission(cell, smear, infil, li, blast, temp)

remiss	cell	smear	infil	li	blast	temp
1	0.8	0.83	0.66	1.9	1.1	0.996
1	0.9	0.36	0.32	1.4	0.74	0.992
0	0.8	0.88	0.7	0.8	0.176	0.982
0	1	0.87	0.87	0.7	1.053	0.986
1	0.9	0.75	0.68	1.3	0.519	0.98
0	1	0.65	0.65	0.6	0.519	0.982
1	0.95	0.97	0.92	1	1.23	0.992
0	0.95	0.87	0.83	1.9	1.354	1.02

Remission data (Lee, 1974)



remiss	cell	smear	infil	li	blast	temp	\hat{p} t3\$fitted.values
1	0.8	0.83	0.66	1.9	1.1	0.996	0.722648915
1	0.9	0.36	0.32	1.4	0.74	0.992	0.578739122
0	0.8	0.88	0.7	0.8	0.176	0.982	0.104598953
0	1	0.87	0.87	0.7	1.053	0.986	0.282577342
1	0.9	0.75	0.68	1.3	0.519	0.98	0.714180403
0	1	0.65	0.65	0.6	0.519	0.982	0.270886837
1	0.95	0.97	0.92	1	1.23	0.992	0.321555392

● 로지스틱 회귀모형

- ☑ 로지스틱 회귀분석(logistic regression)은 종속변수가 범주형인 경우
타겟변수가 2개의 범주(양성/음성, 불량/양품 등) 혹은 3개 이상의 범주일때 사용
3개 범주 이상의 경우 서열형 데이터(ordinal data), 명목형 데이터(nominal data)에 따라
다른 모형이 사용됨

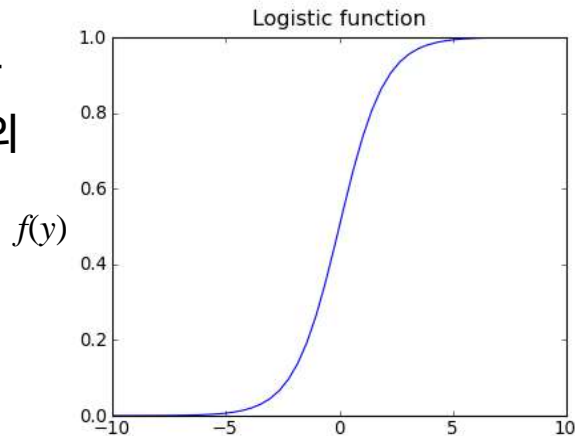
$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) \quad \text{logit}(p) = \beta_0 + \beta_1 x$$

- ☑ 회귀계수 β_1 의 의미는 선형회귀 모형에서와는 다름
 β_1 는 X가 한 단위 증가할 때 $\text{logit}(p)$, 즉 승산비의 로그값의 증가분을 말하므로 승산비가 e^{β_1} 배로 증가함을 의미함

로지스틱 회귀모형

- ☑ Y 가 (0/1, cancer/no cancer, present/absent) 등의 값을 취하는 경우, 다음과 같은 로지스틱 함수가 독립변수들과 Y간의 관계를 설명하기 위해 사용됨

$$f(y) = \frac{1}{1 + e^{-y}}$$



- ☑ 로짓모형은 $Y = (0, 1)$

$$P_i = \Pr\{Y_i = 1 \mid X_1, \dots, X_k\} \quad P_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k)}}$$

$$\text{OR}_{X_1=1 \text{ vs } X_1=0} = \frac{\hat{\text{Odds}}(Y = 1 \mid X_1 = 1, X_2, \dots, X_k)}{\hat{\text{Odds}}(Y = 1 \mid X_1 = 0, X_2, \dots, X_k)} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_k X_k}}{e^{\hat{\beta}_0 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_k X_k}} = e^{\hat{\beta}_1}$$

회귀계수의 해석

로지스틱 회귀모형 - 예제

✓ 로지스틱 회귀모형 : y는 binomial variable, logit function 선택

```
#logistic regression (full model)
t1<-glm(remiss+cell+smear+infil+li+blast+temp, data=re,family=binomial(logit))
summary(t1)

cor(re)
```

```
> t1<-glm(remiss+cell+smear+infil+li+blast+temp, data=re,family=binomial(logit))
> summary(t1)

Call:
glm(formula = remiss ~ cell + smear + infil + li + blast + temp,
    family = binomial(logit), data = re)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.95165  -0.66491  -0.04372   0.74304   1.67069

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  58.0385    71.2364   0.815   0.4152
cell         24.6615    47.8377   0.516   0.6062
smear        19.2936    57.9500   0.333   0.7392
infil       -19.6013    61.6815  -0.318   0.7507
li           3.8960     2.3371   1.667   0.0955
blast         0.1511     2.2786   0.066   0.9471
temp        -87.4339    67.5735  -1.294   0.1957

---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 34.372  on 26  degrees of freedom
Residual deviance: 21.751  on 20  degrees of freedom
AIC: 35.751
```

check p-value
and correlation

```
> cor(re)
```

	remiss	cell	smear	infil
remiss	1.0000000	0.2645288	0.1994882	0.26399695
cell	0.2645288	1.0000000	0.2917881	0.60707674
smear	0.1994882	0.2917881	1.0000000	0.92970107
infil	0.2639970	0.6070767	0.9297011	1.00000000
li	0.5419818	0.1902352	0.3174573	0.32114358
blast	0.3613662	0.4387766	0.6113205	0.69445195
temp	-0.1562395	0.1081586	-0.1124464	-0.04444844

	li	blast	temp
remiss	0.54198183	0.3613662	-0.15623948
cell	0.19023518	0.4387766	0.10815860
smear	0.31745727	0.6113205	-0.11244639
infil	0.32114358	0.6944519	-0.04444844
li	1.00000000	0.6036873	-0.05477238
blast	0.60368727	1.0000000	0.21639325
temp	-0.05477238	0.2163933	1.00000000

blast, infil 제거 후 다시 수행

로지스틱 회귀모형 - 예제

✓ 로지스틱 회귀모형의 평가척도 : $-2\text{Log}(\text{Deviance})$, AIC, likelihood ratio test(G^2)

```
> t2<-glm(remiss~cell+smear+li+temp, data=re,family=binomial(logit))
> summary(t2)
```

Call:
glm(formula = remiss ~ cell + smear + li + temp, family = binomial(logit),
data = re)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.87933	-0.66813	-0.07052	0.78408	1.72472

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	70.0994	58.7634	1.193	0.2329
cell	9.8507	7.8263	1.259	0.2082
smear	0.9124	2.9600	0.308	0.7579
li	3.9052	1.8167	2.150	0.0316 *
temp	-85.4447	64.2143	-1.331	0.1833

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 34.372 on 26 degrees of freedom
Residual deviance: 21.858 on 22 degrees of freedom
AIC: 31.858

로지스틱 회귀모형 - 예제

```
> t3<-glm(remiss~cell+li+temp, data=re,family=binomia
> summary(t3)
```

Call:
glm(formula = remiss ~ cell + li + temp, family = bin
data = re)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.02043	-0.66313	-0.08323	0.81282	1.65887

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	67.634	56.888	1.189	0.2345
cell	9.652	7.751	1.245	0.2130
li	3.867	1.778	2.175	0.0297 *
temp	-82.074	61.712	-1.330	0.1835

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

(Dispersion parameter for binomial family taken to be

Null deviance:	34.372	on 26	degrees of freedom
Residual deviance:	21.953	on 23	degrees of freedom

AIC: 29.953

1 Logistic equation

$$\text{logit}(p) = 67.63 + 9.65 \text{Cell} + 3.87 \text{Li} - 82.07 \text{Temp}$$

2

$$p_i = \frac{e^{-(67.63+9.65\text{Cell}+3.87\text{Li}-82.07\text{Temp})}}{1 + e^{-(67.63+9.65\text{Cell}+3.87\text{Li}-82.07\text{Temp})}}$$

3 overall significant

4 li 1단위 증가하면 remission 될 확률 $\exp(3.867)=47.79$

앞의 모형(smear 포함, AIC=31.86)보다 AIC값이 낮음

로지스틱 회귀모형 - 예제

☑ 주어진 데이터에 대한 로지스틱모형 확률값 계산 예제

✧ 추정회귀식 $\text{logit}(P)$ 에서 Cell=1, Li=1.2, Temp=.99 이면,

$$\text{logit}(p) = 67.63 + 9.65 * 1 + 3.87 * 1.2 - 82.07 * .99$$

$$\text{✧ } p = \frac{\exp(0.68)}{\exp(0.68) + 1} = .6637$$


즉 remission 될 확률은 0.6637.

=> 임계치 (1과 0으로 할당하는)에 따라 이 관측치는 1 혹은 0으로 분류됨.

로지스틱 회귀모형 - 예제

✓ 예측확률값 출력 : 원래 데이터 + 예측확률값

```
# output data with predicted probability  
dat1_pred<-cbind(re,t3$fitted.values)  
write.table(dat1_pred,file="dat1_pred.csv", row.names=FALSE, sep=",", na=" ")
```



remiss	cell	smear	infil	li	blast	temp	t3\$fitted.values
1	0.8	0.83	0.66	1.9	1.1	0.996	0.722648915
1	0.9	0.36	0.32	1.4	0.74	0.992	0.578739122
0	0.8	0.88	0.7	0.8	0.176	0.982	0.104598953
0	1	0.87	0.87	0.7	1.053	0.986	0.282577342
1	0.9	0.75	0.68	1.3	0.519	0.98	0.714180403
0	1	0.65	0.65	0.6	0.519	0.982	0.270886837
1	0.95	0.97	0.92	1	1.23	0.992	0.321555392
0	0.95	0.87	0.83	1.9	1.354	1.02	0.607231948
0	1	0.45	0.45	0.8	0.322	0.999	0.166316409