Wk14-3 : 로지스틱 회귀분석
(Logistic Regression)

1. 로지스틱 회귀모형

Y : Remiss (0, 1)

6 explanatory variables : risk factor related cancer remission (Cell, Smear, infill, 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)

$\hat{p}$

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

1. 로지스틱 회귀모형

- 로지스틱 회귀분석 (logistic regression)은 종속변수가 범주형인 경우 사용되는데 2개의 범주 (양성/음성, 불량/양품 등) 혹은 3개 이상의 범주를 다루기도 한다. 3개 범주 이상의 경우 서열형 데이터 (ordinal data)인 경우와 명목형 데이터 (nominal data)에 따라 다른 모형이 사용된다.

$$\log it(p) = \ln(\frac{p}{1-p}) \qquad \log it(p) = \beta_0 + \beta_1 X$$

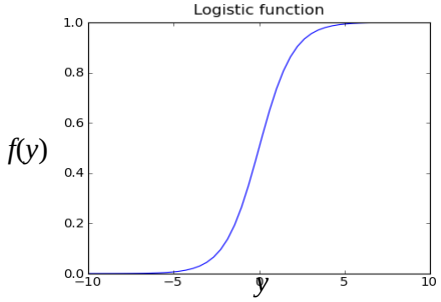
- 회귀계수 β_1의 의미는 선형회귀모형에서와는 다르다. β_1는 X가 한 단위 증가할 때 logit(P), 즉 승산비의 로그값의 증가분을 말하므로 승산비가 배로 증가함을 의미한다.

14-3 로지스틱 회귀분석

1. 로지스틱 회귀모형

- Y 가 (0/1, cancer/no cancer, present/absent) 등의 값을 취하는 경우, 다음과 같은 로지스틱함수가 독립변수들과 Y간의 관계를 설명하기 위해 사용된다.

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



- 로짓모형은 Y= (0,1)

$$P_i = \Pr\{Y_i = 1 \mid X_1, \ldots, X_k\}$$

$$P_i = \frac{1}{1+e[-(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)}$$

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

회귀계수의 해석

2. 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(p1)
          remiss      cell      smear       infil          li     blast
remiss 1.0000000 0.2645288  0.1994882  0.26399695  0.54198183 0.3613662
cell   0.2645288 1.0000000  0.2917881  0.60707674  0.19023518 0.4387766
smear  0.1994882 0.2917881  1.0000000  0.92970107  0.31745727 0.6113205
infil  0.2639970 0.6070767  0.9297011  1.00000000  0.32114358 0.6944519
li     0.5419818 0.1902352  0.3174573  0.32114358  1.00000000 0.6036873
blast  0.3613662 0.4387766  0.6113205  0.69445195  0.60368727 1.0000000
temp  -0.1562395 0.1081586 -0.1124464 -0.04444844 -0.05477238 0.2163933
```

blast, infill 제거후 다시 수행

2. 로지스틱 회귀모형 - 예제

- 로지스틱 회귀모형의 평가척도 : -2Log (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
```

2. 로지스틱 회귀모형 - 예제

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

Call:
glm(formula = remiss ~ cell + li + temp, family = binomial(logit),
    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 '.' 0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 34.372  on 26  degrees of freedom
Residual deviance: 21.953  on 23  degrees of freedom
AIC: 29.953
```

1. logistic equation

: logit(p)= $67.63 + 9.65Cell + 3.87Li - 82.07Temp$

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

3. Overall significant !!

4. Li 1단위 증가하면 remission 될 확률 exp(3.867)=47.79.

14-3 로지스틱 회귀분석

2. 로지스틱 회귀모형 - 예제

[예제] 계속

위의 결과에 대한 추정회귀식은 다음과 같다.

logit(P)= 67.64 + 9.65Cell + 3.87 li – 82.07 temp

Cell=1 이고 li=1.2, temp=.99 이면,

logit(P)= 67.64 + 9.65*1 + 3.87 *1.2 – 82.07 *.99 =0.68

$p = \frac{\exp(0.68)}{\exp(0.68)+1} = .6637$ **즉 remission 될 확률은 0.6637이라고 할 수 있다.**

https://stats.idre.ucla.edu/r/dae/logit-regression/

2. 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