로지스틱 회귀 분석 실습

- 1986년 우주왕복선 챌린저호 사고

데이터 셋 개요

- 1986년 챌린저호 발사 후 73초만에 폭발 후, 대서양 추락 7명의 승무원이 전원 사망
- 원인 : 고체연료 부스터인 부품인 O링이 망가졌다.
- O링이 셔틀 출발 시처럼 낮은 온도에서 작동하도록 설계되지 않았다.

로지스틱 회귀

- 성공-실패 범주형 y변수와 수량형 설명 변수를 가진 데이터는 전통적인 선형 모형으로 다룰 수 없다.
- 데이터 셋 URL : https://raw.githubusercontent.com/stedy/Machine-Learning-with-R-datasets/master/challenger.csv (https://raw.githubusercontent.com/stedy/Machine-Learning-with-R-datasets/master/challenger.csv)

In [3]:
```
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

01. 데이터 불러오기

In [4]:
```
chall <- read.csv("https://raw.githubusercontent.com/stedy/Machine-Learning-with-R-datasets/master/challenger.csv")
chall <- tbl_df(chall)
glimpse(chall)
```

```
Observations: 23
Variables: 5
$ o_ring_ct   <int> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6...
$ distress_ct <int> 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 2, 0, 0, 0, 0, 0...
$ temperature <int> 66, 70, 69, 68, 67, 72, 73, 70, 57, 63, 70, 78, 67, 53,...
$ pressure    <int> 50, 50, 50, 50, 50, 50, 100, 100, 200, 200, 200, 200, 2...
$ launch_id   <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, ...
```

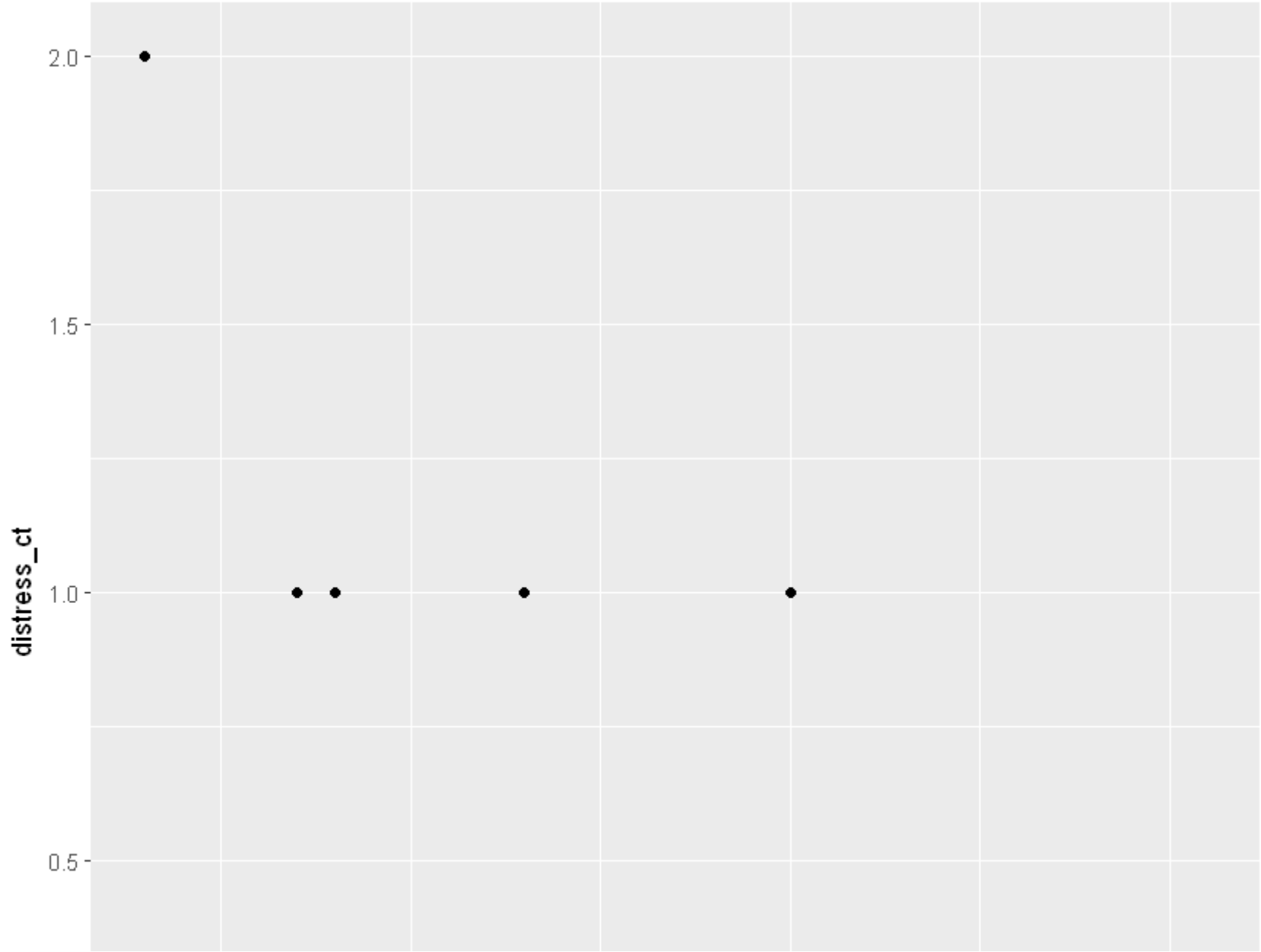
- 여러가지 변수 중 temp, distress_ct 만 고려한다.
- temperature : 온도, distress_ct : 6개의 링 중에서 몇 개의 O링이 실패했는가?

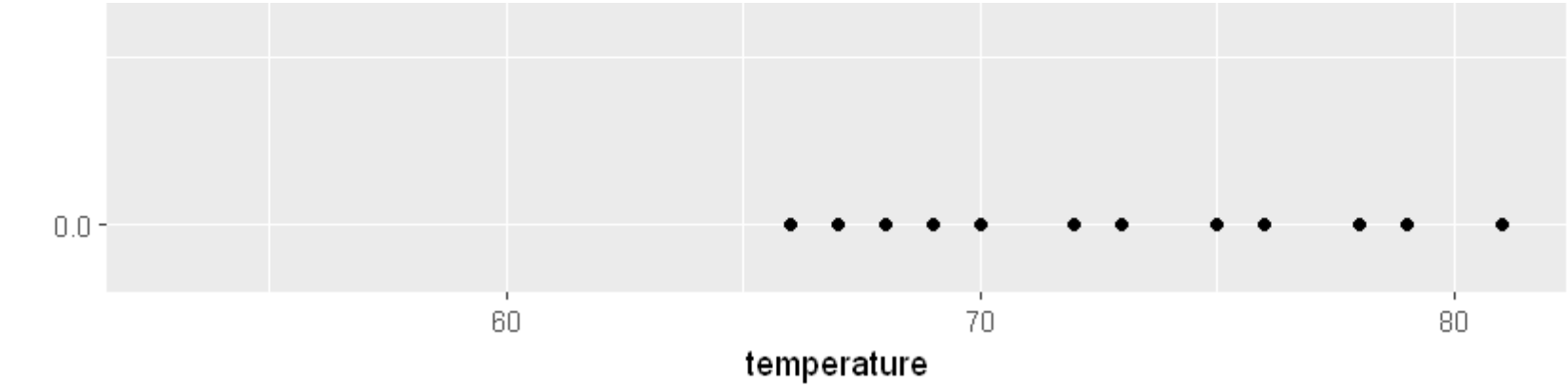
In [9]:
```
summary(chall)
```

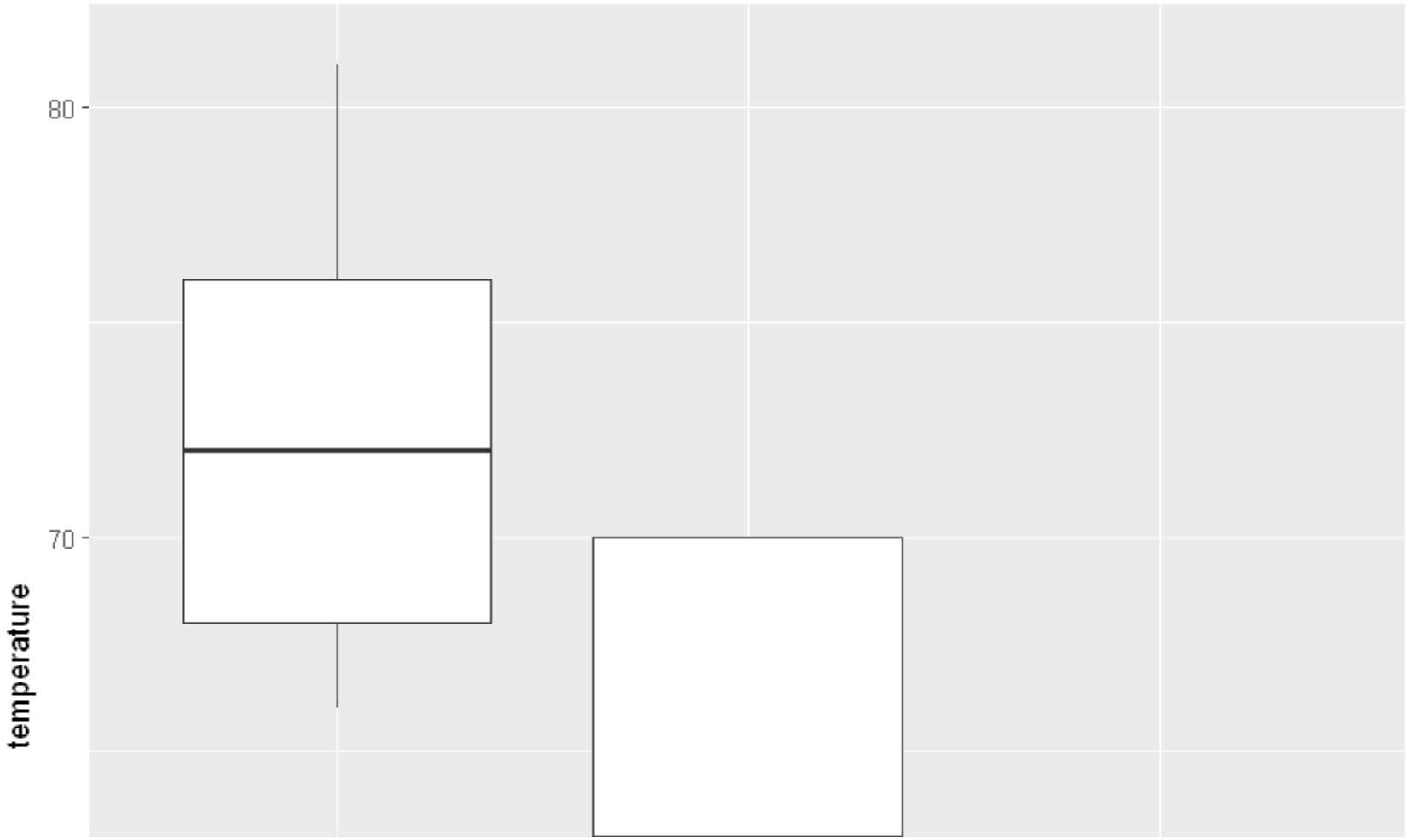
```
   o_ring_ct  distress_ct      temperature      pressure       launch_id    
 Min.   :6   Min.   :0.0000   Min.   :53.00   Min.   : 50.0   Min.   : 1.0  
 1st Qu.:6   1st Qu.:0.0000   1st Qu.:67.00   1st Qu.: 75.0   1st Qu.: 6.5  
 Median :6   Median :0.0000   Median :70.00   Median :200.0   Median :12.0  
 Mean   :6   Mean   :0.3043   Mean   :69.57   Mean   :152.2   Mean   :12.0  
 3rd Qu.:6   3rd Qu.:0.5000   3rd Qu.:75.00   3rd Qu.:200.0   3rd Qu.:17.5  
 Max.   :6   Max.   :2.0000   Max.   :81.00   Max.   :200.0   Max.   :23.0  
```

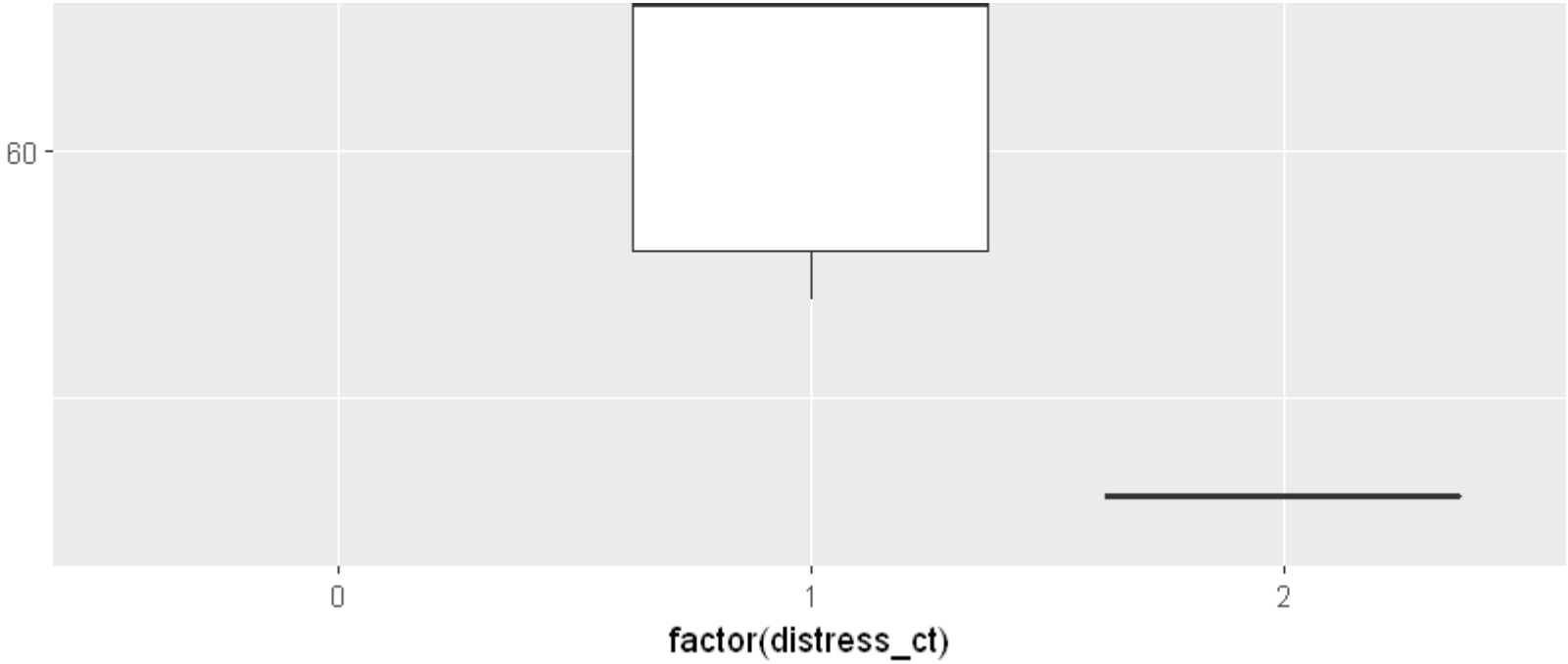
In [6]:
```
library(ggplot2)
```

In [7]:
```
chall %>% ggplot(aes(temperature, distress_ct)) + geom_point()
chall %>% ggplot(aes(factor(distress_ct), temperature)) + geom_boxplot()
```




0.0
60
70
80
temperature


80
70
temperature



02. glm() 함수를 이용하여 모형 만들기

- s='성공횟수', a='시도횟수'
- glm(, family="binomial")
- TRUE, FALSE 이면 0=FALSE(실패) 1=TRUE(성공)으로 간주
- 2레벨 이상 팩터(factor)변수는 첫번째 레벨은 '실패', 나머지 레벨은 '성공'으로 간주
- 반응변수가 2차원 매트릭스이면 첫 열은 '성공' 횟수 s, 두번째 열은 '실패'횟수 a-s를 나타낸다.

In [11]:
```
### 여기에서 o_ring의 '실패'를 성공으로 정의함.
attach(chall)
cbind(distress_ct, o_ring_ct - distress_ct)
detach(chall)
```

```
The following object is masked from package:datasets:

    pressure
```

distress_ct	
0	6
1	5
0	6
0	6
0	6
0	6
0	6
0	6
0	6
1	5
1	5
1	5
0	6
0	6
2	4
0	6
0	6
0	6
0	6
0	6

distress_ct	
0	6
0	6
0	6
1	5

In [13]:
```
(glm_model <- glm(cbind(distress_ct, o_ring_ct - distress_ct) ~
                  temperature, data=chall, family='binomial'))
```

```
Call:  glm(formula = cbind(distress_ct, o_ring_ct - distress_ct) ~ temperature,
    family = "binomial", data = chall)

Coefficients:
(Intercept)  temperature
     8.8169      -0.1795

Degrees of Freedom: 22 Total (i.e. Null);  21 Residual
Null Deviance:      20.71
Residual Deviance: 9.527        AIC: 24.87
```

In [15]:
```
summary(glm_model)
```

```
Call:
glm(formula = cbind(distress_ct, o_ring_ct - distress_ct) ~ temperature,
    family = "binomial", data = chall)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7526  -0.5533  -0.3388  -0.1901   1.5388

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  8.81692    3.60697   2.444  0.01451 *
temperature -0.17949    0.05822  -3.083  0.00205 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 20.706  on 22  degrees of freedom
Residual deviance:  9.527  on 21  degrees of freedom
AIC: 24.865

Number of Fisher Scoring iterations: 6
```

- Number of Fisher Scoring iterations : 6 => 이 모델은 해의 공식이 존재하지 않으므로 수치적인 방법으로 답을 점진적으로 찾아낸다. 이 반복횟수가 6번이다.
- temperature 의 p-value 는 0.002로 극단적인 확률이므로
- 'o링의 실패에 영향이 없다'는 귀무가설은 아마 사실이 아니다.
- temperature -0.17949는 온도가 1도 상승할 때, 우측 값은 0.179만큼 감소한다.
- 마지막 두줄은 모형의 적합도(goodness of fit)를 나타낸다.

```
    Null deviance: 20.706  on 22  degrees of freedom
    Residual deviance:  9.527  on 21  degrees of freedom
```

- deviance 편차이다. 숫자가 높을 수록 모형이 적합하지 않다.
- Null deviance, Residual Deviance --> https://goo.gl/kVotuV (https://goo.gl/kVotuV) 참조(wiki)

- Null deviance와 Residual Deviance가 충분히 줄었다면 이 모형은 적합하다고 판단한다.
- 20.7 - 9.52 = 11.2 (자유도 1인 카이제곱 분포에서 이 값은 아주 아주 나오기 힘든 값.

In [17]:
```
### 카이제곱 분포에서 11.2, 자유도 1이 나올 확률
1-pchisq(11.2,1)
```

0.000817973319994447

- AIC 값 : 2k - 2ln(L)로 주어진다.
- k는 모형 모수의 개수
- L은 주어진 모형으로 최대화한 우도(https://goo.gl/w42U1j (https://goo.gl/w42U1j) 참조)

예측

In [18]:
```
predict(glm_model, data.frame(temperature=30))
```

1: 3.43215903839514

- 내부적으로 predict는 predict.glm()이 호출됨.
- predict.glm() 은 type=c('link', 'response', 'terms') 옵션
- 기본 link는 선형 예측값을 출력한다. 8.82 - 0.179 * 30 = 3.43
- 선형 예측값이 아닌 확률값을 얻으려면 predict.glm()에서 type='response' 옵션을 사용

In [19]:
```
exp(3.45) / (exp(3.45) + 1)
```

0.969231140642852

In [20]:
```
predict(glm_model, data.frame(temperature=30), type='response')
```

1: 0.968694607674951

glm 모형의 일반화

GLM 모형의 일반화

패밀리	디폴트 링크함수	적용 예
binomial	link='logit'	성공-실패 반응변수
gaussian	link='identity'	선형 모형이다! lm( ) 함수와 같다.
Gamma	link='inverse'	양의 값을 가지는 수량형 반응변수
poisson	link='log'	0,1,2, ...값을 가진 '개수'를 나타내는 반응변수 (일일 교통사고 횟수, 단위지역의 연간 지진발생 횟수 등)

- GLM 적절한 활용을 위해 Agresti의
- Nelder & McCullaugh의 등을 참고

In [ ]: