베이즈데이터분석 / 이재용 교수

14강

일반화 선형모형

목차

- 분류에 대한 소개
- 로지스틱모형
- 포아송 회귀모형

목차

- 분류에 대한 소개
- 로지스틱모형
- 포아송 회귀모형

분류(classification)

분류란

반응변수가 범주형 변수(categorical variable)인 회귀분석

분류의 예들

- 응급실에 도착한 환자의 상태 : 심각, 양호
- 온라인 결제 은행은 IP 주소,
  과거 결제기록 등에 기반해 사기인가, 아닌가 판단
- DNA를 기반해 병의 발생을 예측

분류(classification)

분류기(classifier)는

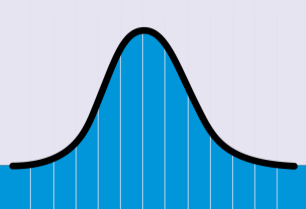
분류 방법론을 말한다.

분류기의 종류

- 로지스틱 회귀분석(logistic regression)
- 선형판별분석(linear discriminant analysis, LDA)
- 최근접이웃방법(K-nearest neighbors, KNN)
- 나무모형(tree models)
- 랜덤숲(random forest)
- 부스팅(boosting)
- 서포트벡터머신(support vector machine) 등이 있다.

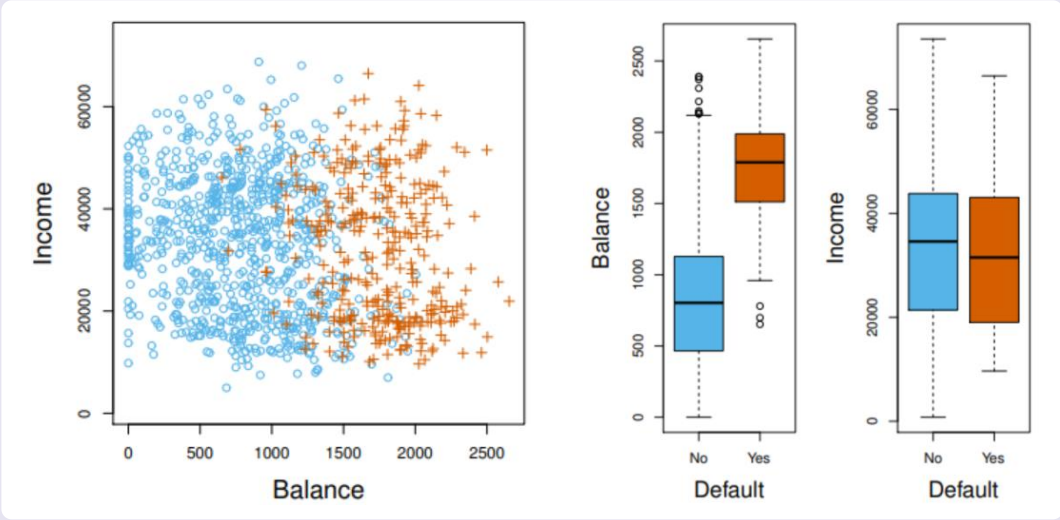
목차

- 분류에 대한 소개
- 로지스틱모형
- 포아송 회귀모형



디폴트 자료

- 모의 자료이다.
- 변수는 default, student, (annual) income, (monthly) balance (잔고)가 있다.



R 코드: 자료 탐색

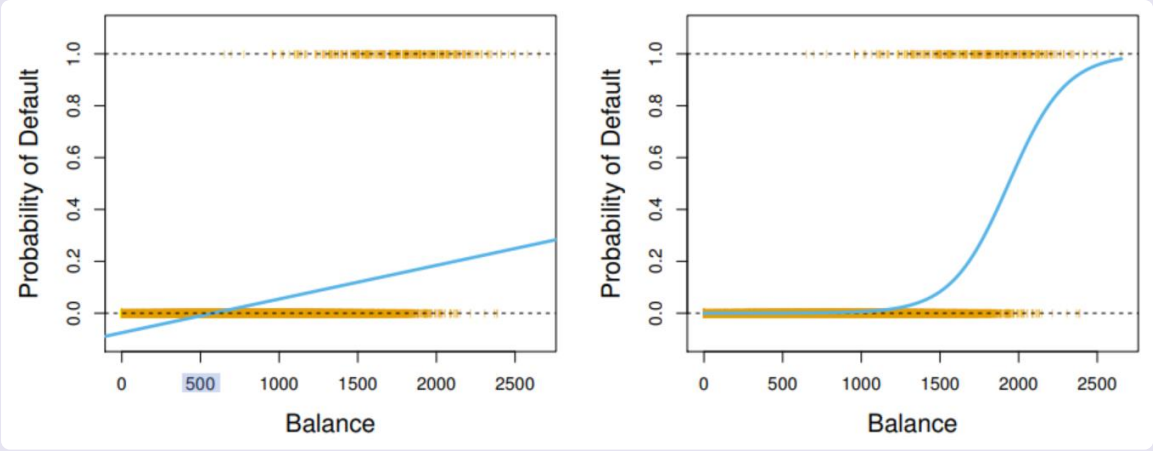
```
summary(Default)
Hmisc::describe(Default)

hist(Default)
ggpairs(Default)
```

로지스틱 회귀모형

$y = 1$인 확률

$$p(x) = logistic(\beta_0 + \beta_1 x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

$$logistic(x) = \frac{e^{x}}{1 + e^{x}}$$



오즈(odds)

$$\frac{p(x)}{1 - p(x)} = e^{\beta_0+\beta_1 x}$$

로짓(logit) 함수

$$logit(p(x)) = \log\frac{p(x)}{1 - p(x)} = \beta_0 + \beta_1 x$$

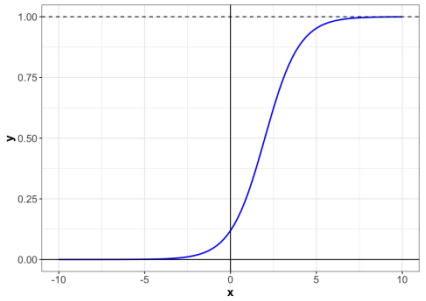
β_0의 의미

- $\mathbb{P}(y = 1|x) = logistic(\beta_0 + \beta_1 x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$

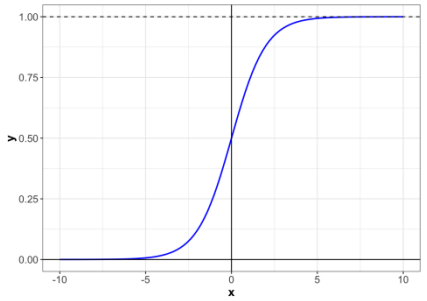
- $x = 0$일 때,

$$\mathbb{P}(y=1|x=0) = \frac{e^{\beta_0+\beta_1\times 0}}{1+e^{\beta_0+\beta_1\times 0}} = \frac{e^{\beta_0}}{1+e^{\beta 0}}, logit(p(0)) = \log\frac{\mathbb{P}(y=1|x=0)}{1-\mathbb{P}(y=1|x=0)} = \beta_0.$$

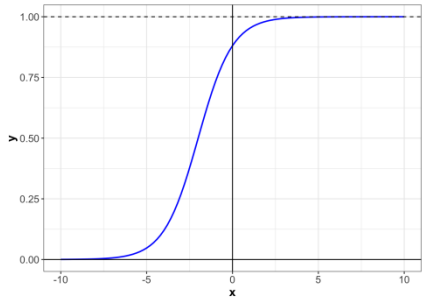
- β_0는 $x = 0$일 때, $\mathbb{P}(y = 1|x = 0)$의 로그-오즈이다.



$\beta_0 < 0$



$\beta_0 = 0$



$\beta_0 > 0$

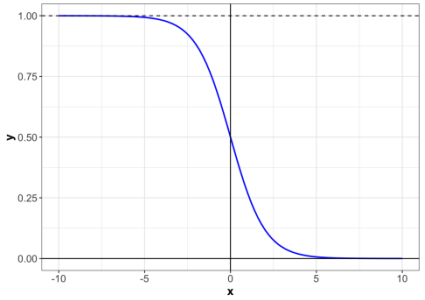
β_1의 의미

- $\mathbb{P}(y = 1|x) = logistic(\beta_0 + \beta_1 x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$
- 로그-오즈의 기울기

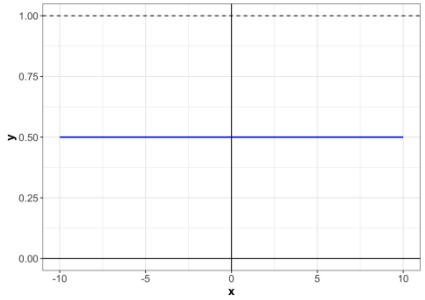
$$\beta_1 = \frac{p(x)\text{의 로그} - \text{오즈의 변화량}}{x\text{의 변화량}}$$

은 x의 값이 변해도 일정하다.

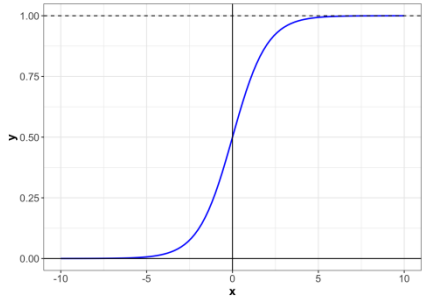
- β_1값의 변화에 따른 $\mathbb{P}(y = 1|x)$의 변화



$\beta_1 < 0$



$\beta_1 = 0$



$\beta_1 > 0$

로지스틱 모형: 사후표본의 추출

스탠에서 모형 서술

```
default.logistic1.code ="
data {
   int<lower=0> n;
   int<lower=0, upper=1> defaults[n];
   vector[n] balance;
}

parameters {
   real alpha;
   real beta;
}

model {
   for(i in 1:n) {
      defaults[i] ~ bernoulli_logit(alpha +
                                    beta * balance[i]);
   }
}
"
```

스탠에서 모형 서술

```
data=list(n=dim(Default)[1],
          defaults=as.integer(as.numeric(Default$default)-1),
          balance=Default$balance)

default.logistic1 = stan(model_code=default.logistic1.code,
          data=data, seed=1234567, chains=1, iter=5000,
          thin=1, cores=4)

print(default.logistic1)
plot(default.logistic1, plotfun="dens")
plot(default.logistic1, plotfun="trace")
plot(default.logistic1, plotfun="ac")
```

로지스틱 모형: 사후표본의 추출

Inference for Stan model: f400a780d2cf3bde1b493f8d70dc458e.
4 chains, each with iter=5000; warmup=2500; thin=1;
post-warmup draws per chain=2500, total post-warmup draws=10000.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
alpha	-10.67	0.01	0.37	-11.39	-10.92	-10.68	-10.41	-9.94	1405	1
beta	0.01	0.00	0.00	0.01	0.01	0.01	0.01	0.01	1443	1
lp__	-799.25	0.02	1.02	-802.01	-799.64	-798.94	-798.52	-798.25	1753	1

Samples were drawn using NUTS(diag_e) at Fri Aug 12 15:26:55 2022.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).

로지스틱 모형: 사후표본의 추출

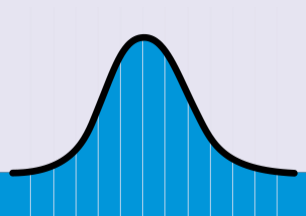
문제

1. 신용카드 잔고가 x 달러일 때,
   파산의 확률 식을 쓰시오.

2. 신용카드 잔고가 $\$1000$인 사람의
   파산 확률은 얼마인가?

목차

- 분류에 대한 소개
- 로지스틱모형
- 포아송 회귀모형



포아송 회귀모형

자료

살모넬라균이 퀴놀린(방부제의 일종, quinoline)의 양에 따라 복귀돌연변이 집락이 몇 개나 생기는지 조사한 자료이다.

Dose	0	10	33	100	333	1000
Plate 1	15	16	16	27	33	20
Plate 2	21	18	33	69	41	42
Plate 3	29	21	33	69	41	42

포아송 회귀모형

모형

기존의 이론은 다음의 모형을 제시한다.

$i = 1, 2, \ldots, 6, j = 1, 2, 3$에 대해,

$$y_{ij} \sim Poi(\mu_i),$$

$$\log \mu_i = \alpha + \beta \log(x_i + 10) + \gamma x_i.$$

혹은

$$\mu_i = e^{\alpha + \beta \log(x_i + 10) + \gamma x_i}.$$

포아송 회귀모형

```
y= t(matrix(c(15, 16, 16, 27, 33, 20, 21, 18, 26, 41, 38, 27, 29, 21, 33, 69, 41, 42),
                nr=3, byrow=TRUE) )
x = c(0,10,33,100,333,1000)

pois.code ="
data {
    real<lower=0> x[6];
    int<lower=0> y[6,3];
}

parameters {
    real alpha;
    real beta;
    real gamma;
}
```

포아송 회귀모형

```
model {
    for(i in 1:6) {
      for(j in 1:3) {
          y[i,j] ~ poisson(exp(alpha+beta*log(x[i] +10) + gamma*x[i]));
      }
    }
}
"


data=list(x=x, y=y)

pois = stan(model_code = pois.code, data=data,
                    seed = 123456789, chains = 1,
                    Iter   = 10000,        thin=10, cores=4)
```

포아송 회귀모형

```
print(pois)
plot(pois, plotfun="dens")
plot(pois, plotfun="trace")
plot(pois, plotfun="ac")
```

포아송 회귀 모형: 사후표본의 추출

Inference for Stan model: af147b31b011c0aaafd1d339dcb7ae64.
1 chains, each with iter=10000; warmup=5000; thin=10;
post-warmup draws per chain=500, total post-warmup draws=500.

	mean	se_mean	sd	2.50%	25%	50%	75%	97.50%	n_eff	Rhat
alpha	2.12	0.01	0.23	1.70	1.96	2.11	2.28	2.58	400	1
beta	0.34	0.00	0.06	0.22	0.30	0.35	0.38	0.45	389	1
gamma	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	373	1
lp__	1290.53	0.07	1.40	1286.57	1290.01	1290.93	1291.51	1291.99	365	1

Samples were drawn using NUTS(diag_e) at Fri Aug 12 15:58:33 2022.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).

포아송 회귀 모형: 사후표본의 추출

문제

1. 퀴놀린의 양(x)에 따른 살모넬라균의 복귀돌연변이 집락의 갯수의 평균을 식으로 써보자.

2. $x = 100$일 때 살모넬라균의 복귀돌연변이 집락의 갯수의 평균은 얼마인가?

다음시간

15강

베이즈 통계와 계층 모형