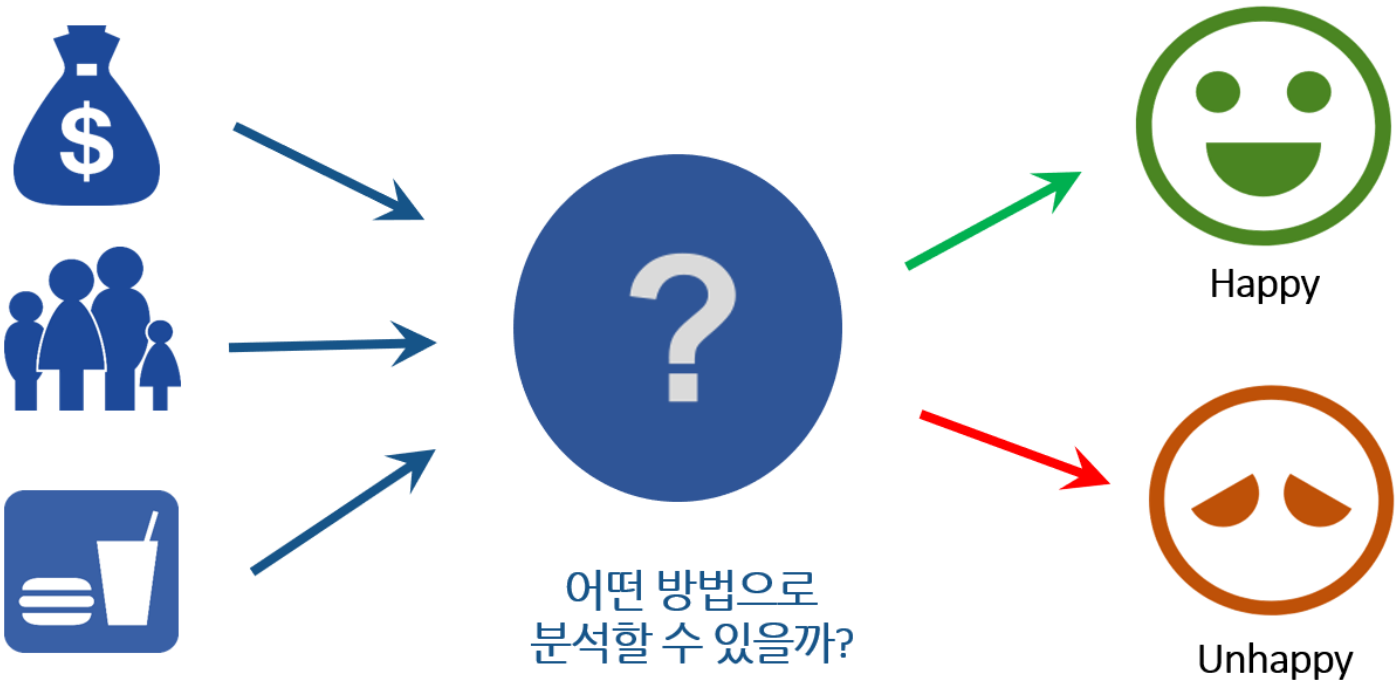
Logistic Regression

Classification



Classification

- 질적자료(Qualitative), 범주형자료(Categorical) :

  eye color $\in \{\text{brown}, \text{blue}, \text{green}\} := \mathcal{C}$

  email $\in \{\text{spame}, \text{ham}\} := \mathcal{C}$.

- 분류(Classification) : 주어진 설명변수 X에 대하여,

  $C(X) \in \mathcal{C}$ 인 함수 C를 찾는 것.

- 때로는 분류 자체 보다 각 범주에 속할 확률에 대한 추정에 관심

Classification

- 반응변수 (Y) : 성공확률이 p인 베르누이 확률분포

$$Y = \begin{cases} 0, & P(Y=0) = 1-p \\ 1, & P(Y=1) = p \end{cases}$$

  - $E(Y) = 0 \times P(Y=0) + 1 \times P(Y=1) = P(Y=1) = p$

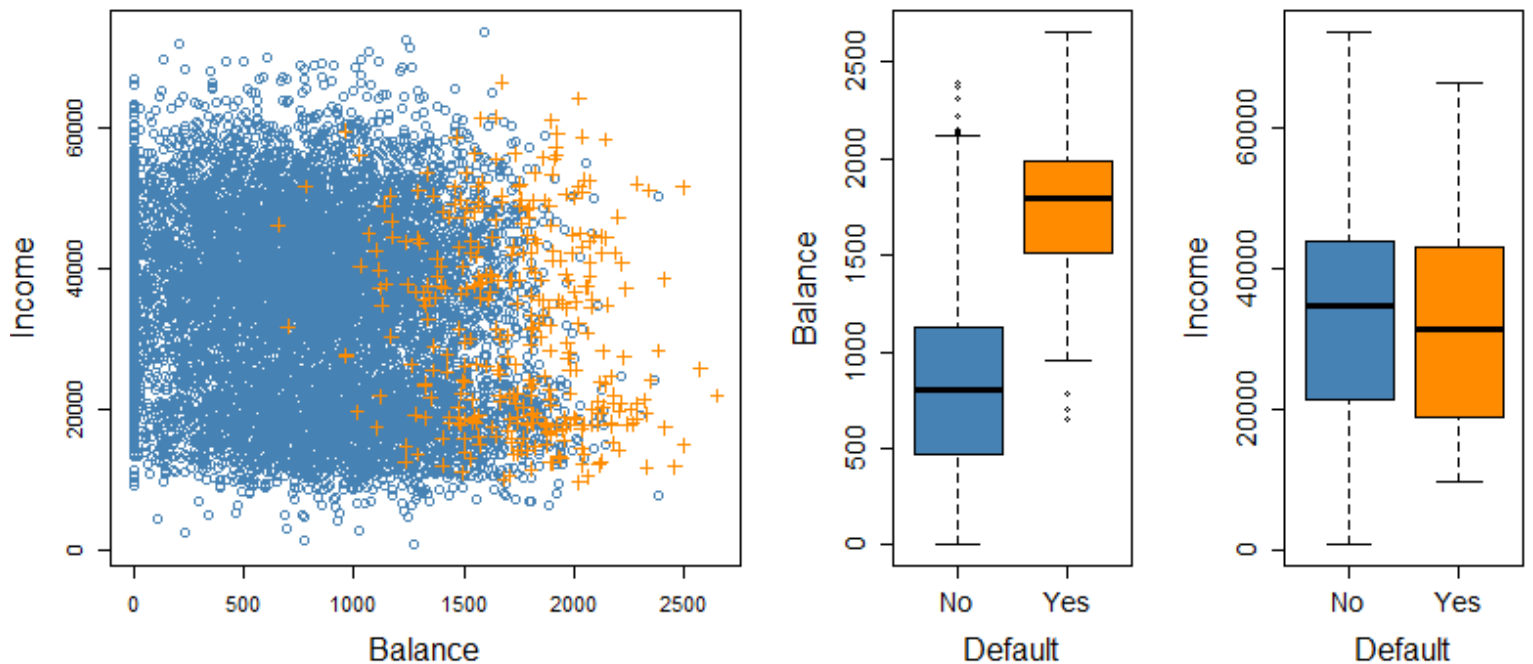
Example

	default	student	balance	income
1	No	No	729.53	44361.63
2	No	Yes	817.18	12106.13
3	No	No	1073.55	31767.14
4	No	No	529.25	35704.49
5	No	No	785.66	38463.50
6	No	Yes	919.59	7491.56
⋮	⋮	⋮	⋮	⋮

Table: Credit Card Default

Credit Card Default

Figure: Scatter plot / Box plot



Linear Regression

- 반응변수 Y : defalut

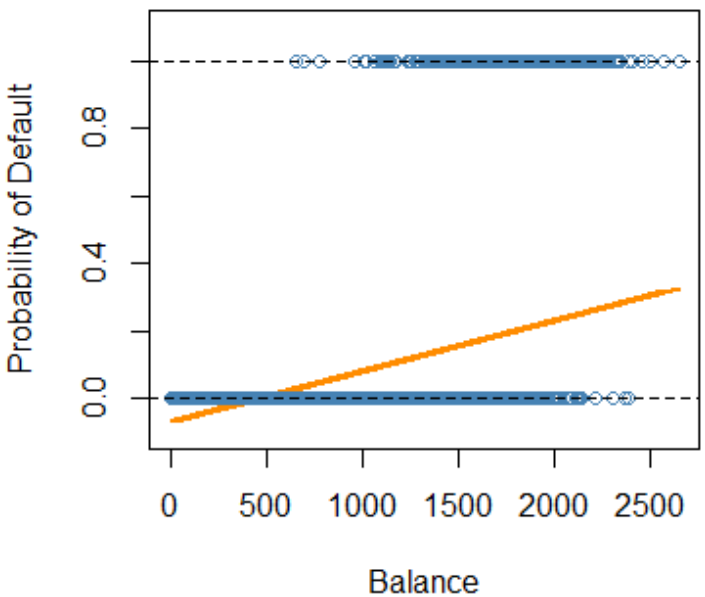
$$Y = \begin{cases} 0 & \text{if No} \\ 1 & \text{if Yes} \end{cases}$$

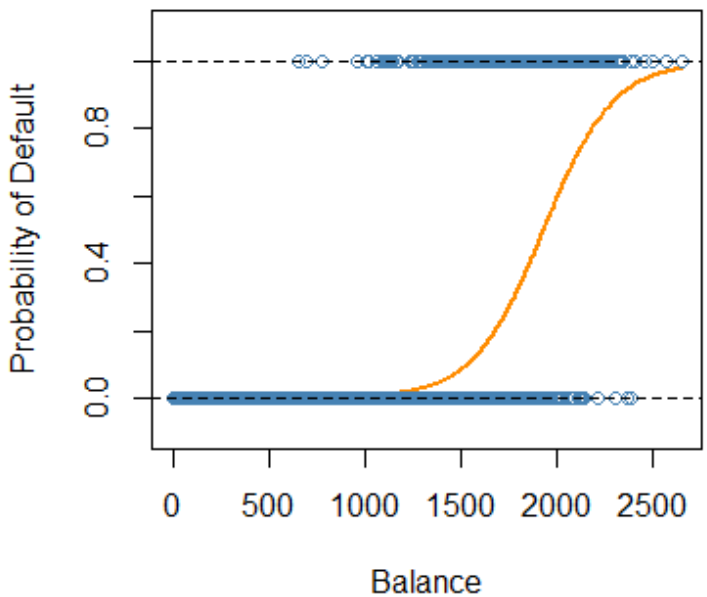
- Linear regression :

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

  - y는 0 또는 1이 아닌 다른 값을 갖을 수 있음
  - $\hat{y} > 0.5 \Rightarrow$ "Yes" ?

선형회귀 vs. 로지스틱 회귀





확률을 이용한 회귀분석

$$E(Y|X = x) = p = \hat{\beta}_0 + \hat{\beta}_1 x$$

- $P(Y = 1|X = x) = p, P(Y = 0|X = x) = 1 - p$
- $0 \leq p \leq 1$
- $\hat{\beta}_0 + \hat{\beta}_1 x$ 의 값은 0과 1사이를 벗어날 수 있음
- 오차항의 분포가 정규분포가 될 수 없음
- 선형회귀모형은 적적하지 않음.

Logistic Regression

- Model : $P(Y = 1|X) = p(X)$

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

- $0 \leq p(X) \leq 1$

- Logit 또는 log odds

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X$$

Logistic Regression

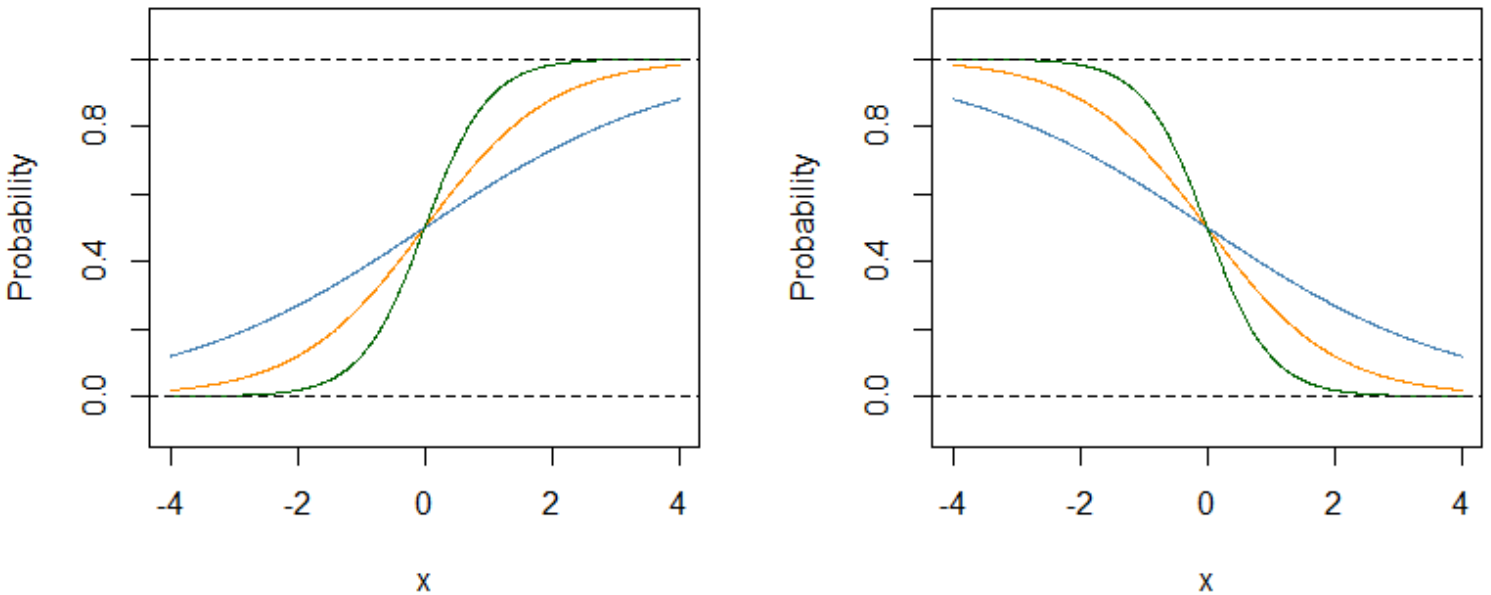


Figure: left : $\beta_1 > 0$ (orange : $\beta_1 = 1$, blue:$\beta_1 = 0.5$, green:$\beta_1 = 2$)

right : $\beta_1 < 0$ (orange:$\beta_1 = -1$, blue:$\beta_1 = -0.5$, green:$\beta_1 = -2$)

계수추정 : Maximum Likelihood

- likelihood function

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}$$

- log-likelihood function

$$\ell(\beta_0, \beta_1) = \sum_{i=1}^{n} \left\{ y_i \log \frac{p(x_i)}{1 - p(x_i)} \right\} + \sum_{i=1}^{n} \log(1 - p(x_i))$$

- Maximum likelihood estimation (MLE)

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{(\beta_0, \beta_1) \in \mathbb{R}^2}{\operatorname{argmin}} \ \ell(\beta_0, \beta_1)$$

Credit Card Default

- 로지스틱 회귀계수 추정값

	Estimate	Std. Error	z value	Pr($>|z|$)
(Intercept)	-10.6513	0.3612	-29.49	< 0.0001
balance	0.0055	0.0002	24.95	< 0.0001

- 추정된 모형 (X : balance)

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times X}}{1+e^{-10.6513+0.0055\times X}}$$

회귀계수의 해석

- balance= \$1000 인 누군가의 default 확률

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.006$$

- balance= \$2000 인 누군가의 default 확률

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 2000}}{1+e^{-10.6513+0.0055\times 2000}} = 0.586$$

Credit Card Default

- 로지스틱 회귀계수 추정값

	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	-3.5041	0.0707	-49.55	< 0.0001
studentYes	0.4049	0.1150	3.52	0.0004

$$\hat{Pr}(\text{defalut=Yes}|\text{student=Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431$$

$$\hat{Pr}(\text{defalut=Yes}|\text{student=No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292$$

Odds

- 오즈 (odds)

$$Odds = \frac{p}{1-p}$$

  - p : 성공 확률, $1-p$: 실패확률
  - 실패할 확률에 대한 성공할 확률의 비

- 오즈비(odds ratio)

$$Odds\ Ratio = \frac{Odds_1}{Odds_2}$$

  - $Odds_1 = \frac{p_1}{1-p_1}$, $Odds_2 = \frac{p_2}{1-p_2}$: 두 그룹의 오즈비

Odds ratio

- 예) X : 이산형 설명변수, $X = 1$ 또는 $X = 0$

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

- $X = 1$ 그룹에서의 Odds

$$Odds_1 = \frac{\hat{Pr}(Y = 1|X = 1)}{1 - \hat{Pr}(Y = 1|X = 1)} = e^{\hat{\beta}_0 + \hat{\beta}_1 X} = e^{\hat{\beta}_0 + \hat{\beta}_1}$$

- $X = 0$ 그룹에서의 Odds

$$Odds_1 = \frac{\hat{Pr}(Y = 1|X = 0)}{1 - \hat{Pr}(Y = 1|X = 0)} = e^{\hat{\beta}_0 + \hat{\beta}_1 X} = e^{\hat{\beta}_0}$$

- $Odds\ Ratio = \dfrac{Odds_1}{Odds_2} = \dfrac{e^{\hat{\beta}_0 + \hat{\beta}_1}}{e^{\hat{\beta}_0}} = e^{\hat{\beta}_1}$

Credit Card Default

- 로지스틱 회귀계수 추정값

	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	-3.5041	0.0707	-49.55	< 0.0001
studentYes	0.4049	0.1150	3.52	0.0004

- $Odds\ Ratio = e^{0.4049} \approx 1.5$
- 학생그룹의 오즈가 학생이 아닌 그룹의 오즈에 비해 1.5배 높다.

설명변수가 여러개인 경우

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	-10.8690	0.4923	-22.08	0.00
balance	0.0057	0.0002	24.74	0.00
income	0.0030	0.0082	0.37	0.7115
studentYes	-0.6468	0.2362	-2.74	0.0062

설명변수가 여러개인 경우