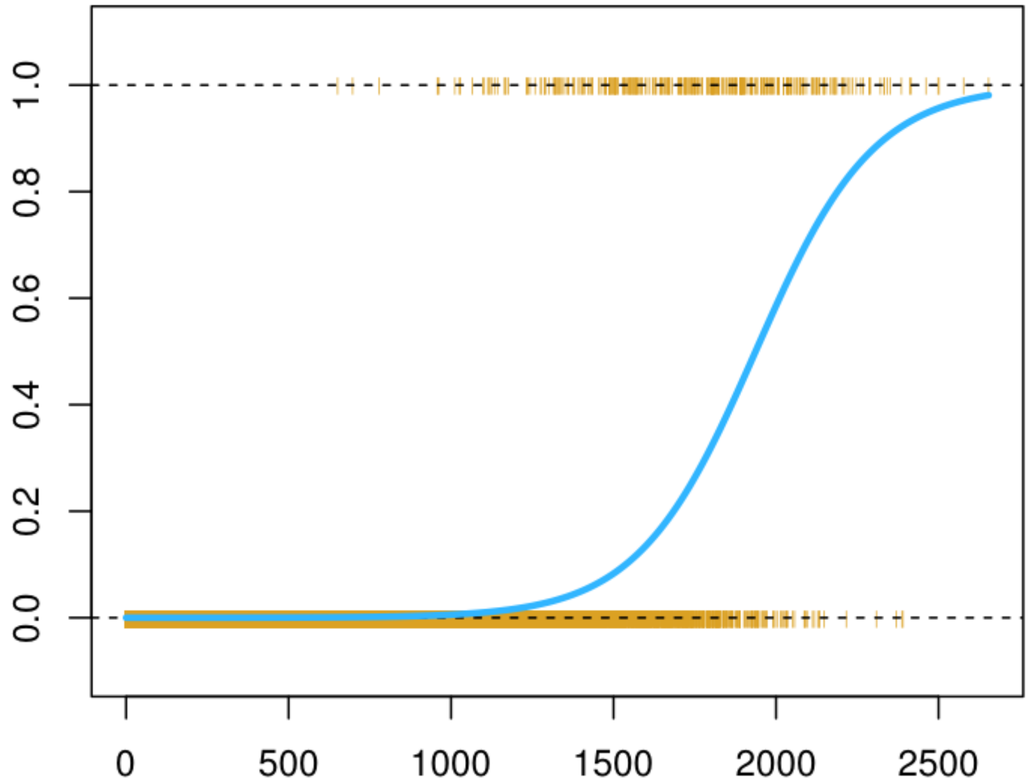

AN
INTRODUCTION
TO
MACHINE
LEARNING
WITH R
DAY 5

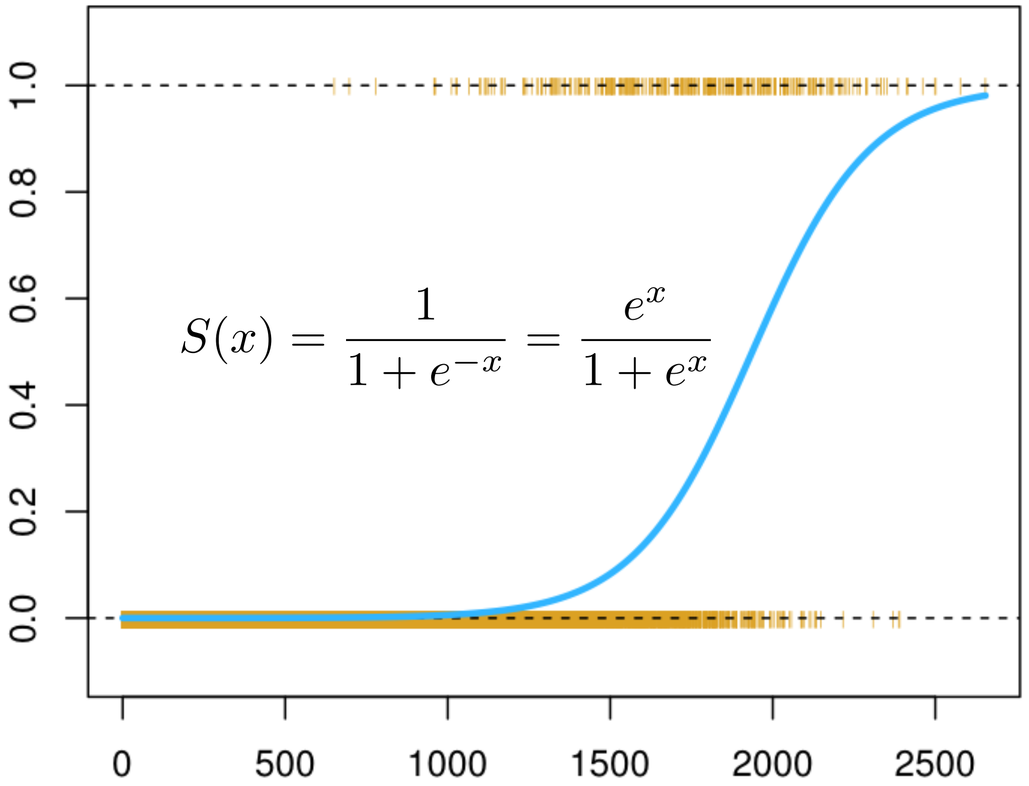

DAY 5
Logistic Regression

Logistic Regression

독립변수와 종속변수 사이의 **관계**를 **사건 발생 가능성**으로 모델링하는 기법



Logistic Regression



독립변수와 종속변수 사이의 **관계**를
사건 발생 가능성으로 모델링하는 기법

categorical data

종속변수의 대상이 **범주형 데이터**일 때
로지스틱 회귀분석을 사용한다.

Logistic Regression

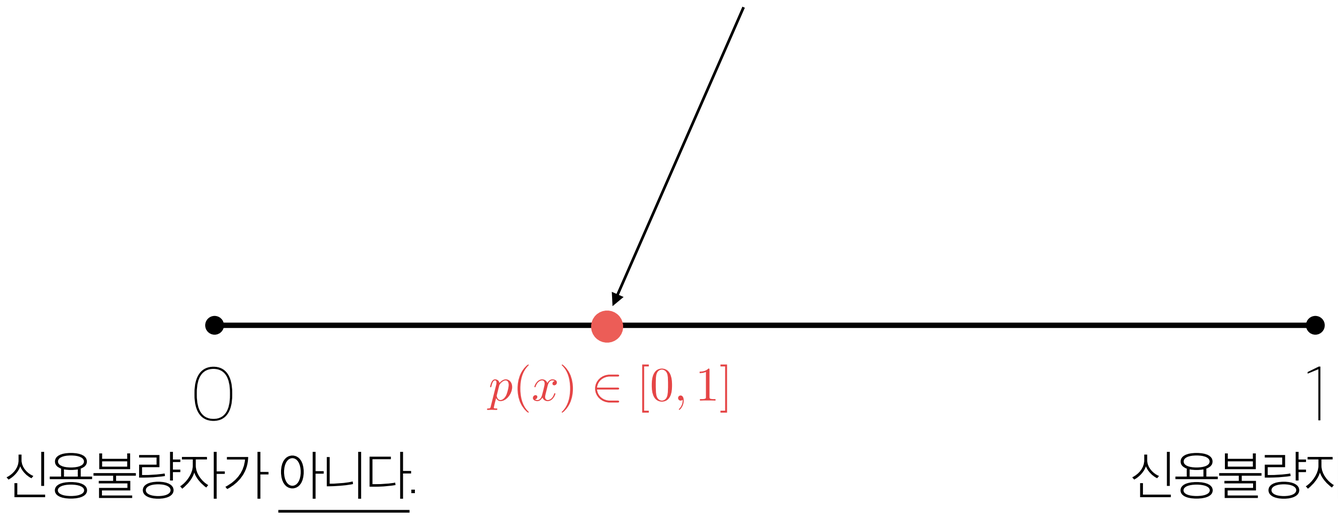
예를 들어,
계좌잔고가 2000달러인 고객이
신용불량자일 확률은 어떻게 될까?

Logistic Regression

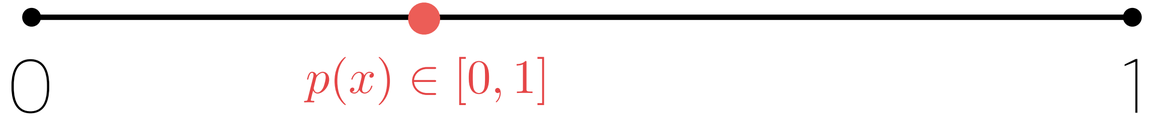
예를 들어,

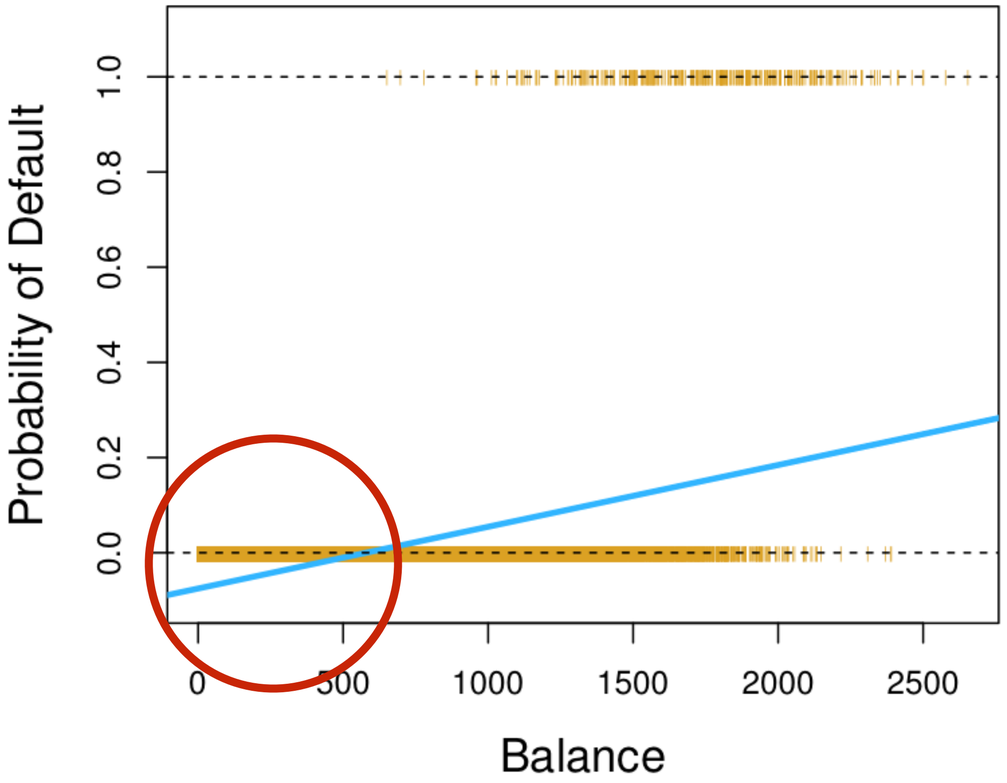
계좌잔고가 2000달러인 고객이

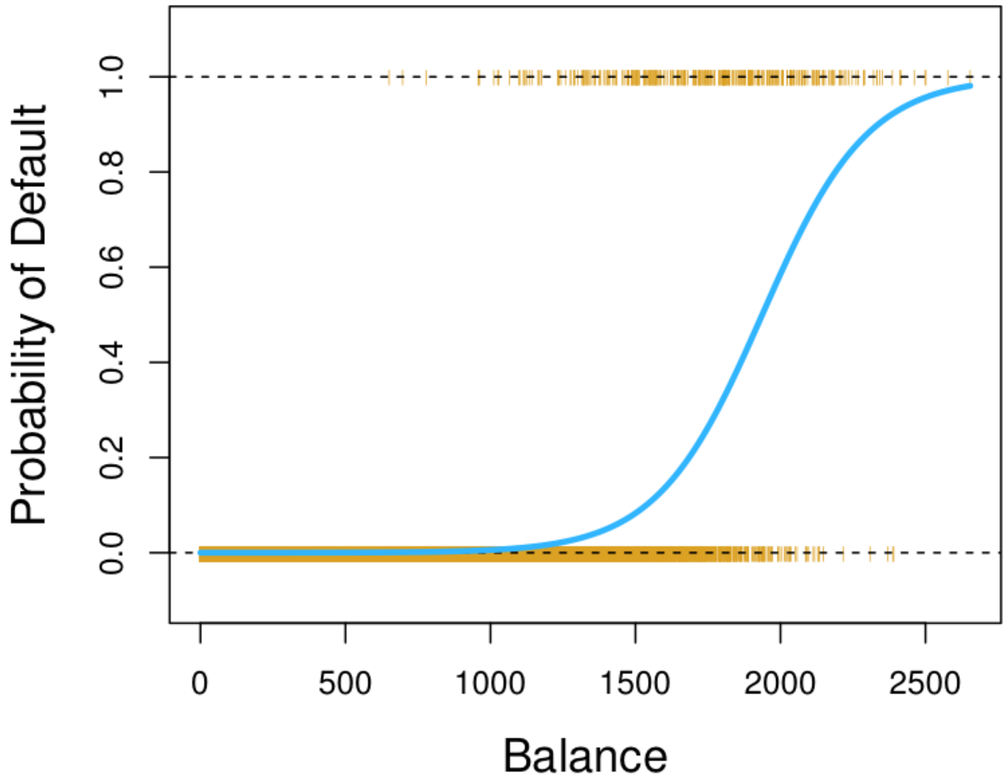
신용불량자일 **확률**은 어떻게 될까?



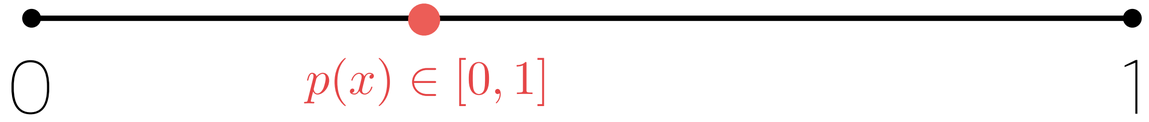
Logistic Regression

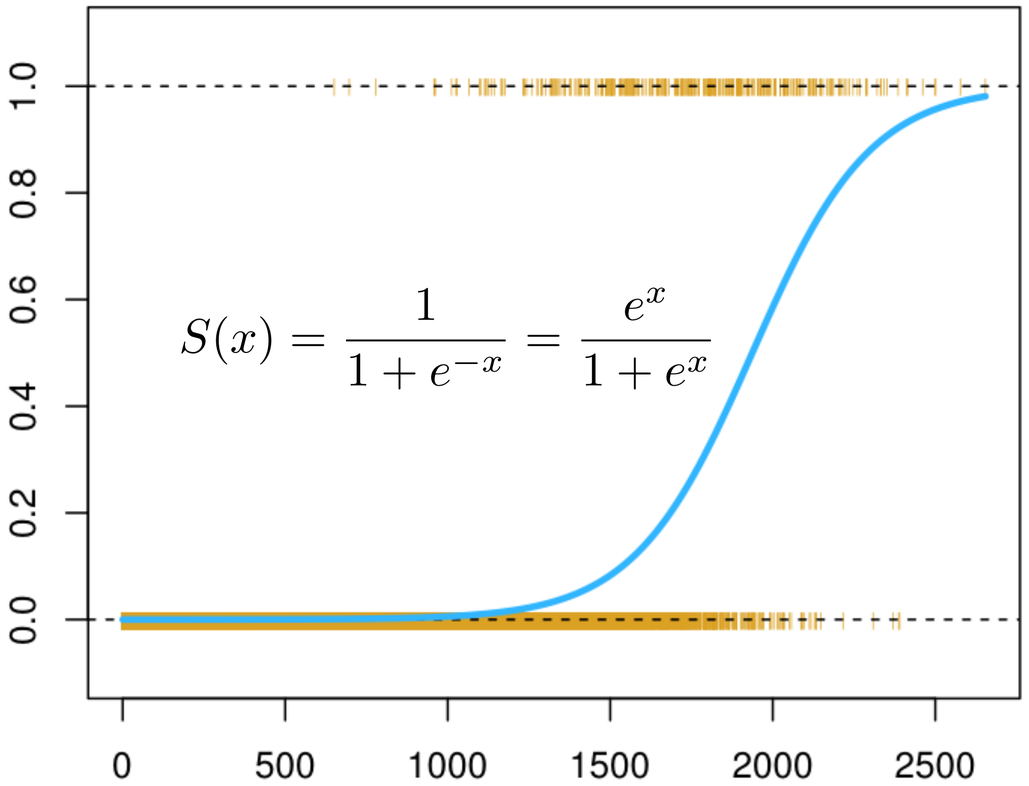






Logistic Regression





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

log-odds

odds $\in [0, \infty)$

Logistic Regression

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

회귀계수를 결정하는 방법은?

최대가능도 방법(Maximum Likelihood Method)


College
Admission
Office
COLLEGE ADMISSIONS
Handicap access

실습

로지스틱 회귀분석

```
Call:
glm(formula = admit ~ gre + gpa + rank, family = "binomial",
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4965  -0.8640  -0.6178   1.1508   1.9943

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.543557   1.360965  -2.604  0.00922 **
gre          0.001331   0.001362   0.978  0.32815
gpa          0.806846   0.402905   2.003  0.04522 *
rank2       -0.484687   0.359375  -1.349  0.17744
rank3       -1.475998   0.409239  -3.607  0.00031 ***
rank4       -1.436535   0.473162  -3.036  0.00240 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 348.59  on 279  degrees of freedom
Residual deviance: 320.05  on 274  degrees of freedom
AIC: 332.05

Number of Fisher Scoring iterations: 4
```

Confusion Matrix

		Actual Condition	
		Actual **Positive**	Actual **Negative**
Predicted Condition	Predicted **Positive**	**True Positive**	**False Positive** (Type I error)
	Predicted **Negative**	**False Negative** (Type II error)	**True Negative**

Confusion Matrix

		Actual Condition	
		Actual **Positive**	Actual **Negative**
Predicted Condition	Predicted **Positive**	**TP**	**FP**
	Predicted **Negative**	**FN**	**TN**

		Actual Condition	
		Actual **Positive**	Actual **Negative**
Predicted Condition	Predicted **Positive**	**TP**	**FP**
	Predicted **Negative**	**FN**	**TN**

Accuracy

Sensitivity

Specificity

Positive Prediction Value

Negative Prediction Value

		Actual Condition	
		Actual **Positive**	Actual **Negative**
Predicted Condition	Predicted **Positive**	**TP**	**FP**
	Predicted **Negative**	**FN**	**TN**

$$Accuracy = (TP + TN) / (P + N) = (TP + TN) / (TP + FP + FN + TN)$$

$$Sensitivity = TP / P = TP / (TP + FN)$$

$$Specificity = TN / N = TN / (FP + TN)$$

$$Positive\ Prediction\ Value = TP / (TP + FP)$$

$$Negative\ Prediction\ Value = TN / (TN + FN)$$

실습

로지스틱 회귀분석

STOCK MARKET PREDICT

```
Call:
glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
    Volume, family = binomial, data = Smarket)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.446  -1.203   1.065   1.145   1.326

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.126000   0.240736  -0.523    0.601
Lag1        -0.073074   0.050167  -1.457    0.145
Lag2        -0.042301   0.050086  -0.845    0.398
Lag3         0.011085   0.049939   0.222    0.824
Lag4         0.009359   0.049974   0.187    0.851
Lag5         0.010313   0.049511   0.208    0.835
Volume       0.135441   0.158360   0.855    0.392

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1731.2  on 1249  degrees of freedom
Residual deviance: 1727.6  on 1243  degrees of freedom
AIC: 1741.6

Number of Fisher Scoring iterations: 3
```

심화

모델 개선하기

[피처 엔지니어링]

FEATURE ENGINEERING

[다중공선성]

MULTICOLLINEARITY

[상호작용항]

INTERACTION TERM

[이상치 제거]

OUTLIER DETECTION

과제

재실 감지 예측

로지스틱 회귀분석을 사용하여
사람이 방에 있는지 없는지를
감지하는 재실 감지를 해보자.

과제

재실 감지 예측

로지스틱 회귀분석을 사용하여
사람이 방에 있는지 없는지를
감지하는 재실 감지를 해보자.

트레이닝 데이터 : `occupancy_train`
테스트 데이터 : `occupancy_test`

1. 아무런 데이터 처리 과정을 거치지 않은 채로 로지스틱 회귀분석을 실시하면 정확도는 **97%**가 나온다.
2. **피처 엔지니어링, 이상치 제거** 등의 방법을 이용해서 정확도가 **98%**가 넘는 모델을 만들자.
3. **Confusion Matrix**를 이용해서 결과를 해석하자.

과제

재실 감지 예측

로지스틱 회귀분석을 사용하여
사람이 방에 있는지 없는지를
감지하는 재실 감지를 해보자.

트레이닝 데이터 : occupancy_train
테스트 데이터 : occupancy_test

```
date                : time year-month-day
                      hour:minute:second
Temperature         : in Celsius
Relative Humidity   : %
Light               : Lux
CO2                 : ppm
Humidity Ratio      : Derived quantity from
                      temperature and relative humidity,
                      in kgwater-vapor/kg-air
Occupancy           : 0 for not occupied,
                      1 for occupied status
```

Handle the Time / Data

```
> head(occupancy_train)
                 date Temperature Humidity Light    CO2 HumidityRatio Occupancy
1 2015-02-04 17:51:00       23.18  27.2720 426.0 721.25   0.004792988         1
2 2015-02-04 17:51:59       23.15  27.2675 429.5 714.00   0.004783441         1
3 2015-02-04 17:53:00       23.15  27.2450 426.0 713.50   0.004779464         1
4 2015-02-04 17:54:00       23.15  27.2000 426.0 708.25   0.004771509         1
5 2015-02-04 17:55:00       23.10  27.2000 426.0 704.50   0.004756993         1
6 2015-02-04 17:55:59       23.10  27.2000 419.0 701.00   0.004756993         1
> str(occupancy_train)
'data.frame':   8143 obs. of  7 variables:
 $ date         : chr  "2015-02-04 17:51:00" "2015-02-04 17:51:59" "2015-02-04 17:53:00" "2015-02-04 17:54:00" ...
 $ Temperature  : num  23.2 23.1 23.1 23.1 23.1 ...
 $ Humidity     : num  27.3 27.3 27.2 27.2 27.2 ...
 $ Light        : num  426 430 426 426 426 ...
 $ CO2          : num  721 714 714 708 704 ...
 $ HumidityRatio: num  0.00479 0.00478 0.00478 0.00477 0.00476 ...
 $ Occupancy    : int  1 1 1 1 1 1 1 1 1 1 ...
```

Handle the Time / Data

```
> occupancy_train$date <- as.POSIXct(occupancy_train$date)
> occupancy_test$date <- as.POSIXct(occupancy_test$date)
> str(occupancy_train)
'data.frame':   8143 obs. of  7 variables:
 $ date         : POSIXct, format: "2015-02-04 17:51:00" "2015-02-04 17:51:59" "2015-02-04 17:53:00" ...
 $ Temperature  : num  23.2 23.1 23.1 23.1 23.1 ...
 $ Humidity     : num  27.3 27.3 27.2 27.2 27.2 ...
 $ Light        : num  426 430 426 426 426 ...
 $ CO2          : num  721 714 714 708 704 ...
 $ HumidityRatio: num  0.00479 0.00478 0.00478 0.00477 0.00476 ...
 $ Occupancy    : int  1 1 1 1 1 1 1 1 1 1 ...
```

Handle the Time / Data

```
library(lubridate)
```

```
weekdays
year
month
hour
minute
second
```


THX :)