

R 기반 의학통계 및 머신러닝

박 승



충북대학교
CHUNGBUK NATIONAL UNIVERSITY

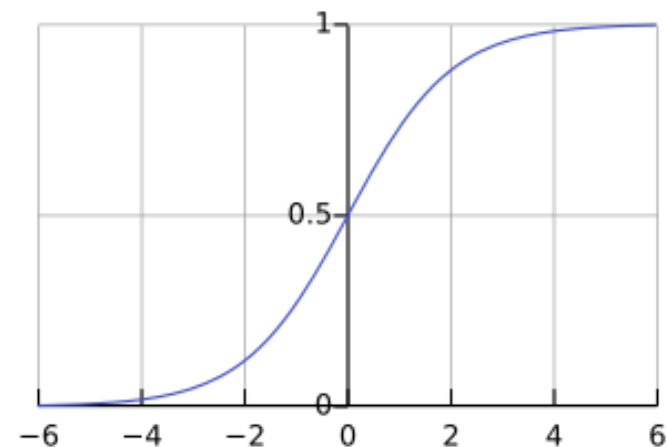
CHAPTER

04

분류 모델을 이용한 질병 예측

■ 로지스틱 회귀분석

- 개요
- 일반적인 회귀분석의 경우는 독립변수와 종속변수가 모두 연속형
- 종속변수가 이분형인 경우 로지스틱 회귀분석을 시행하여 분류 모형을 수립
- 대표적인 머신러닝 방법 중 하나
- 시그모이드 함수를 통해 결과의 범주를 (0,1)로 만든 후
예측값이 0.5보다 작으면 0, 0.5보다 크면 1로 판정
(threshold, 추후 임의대로 조정 가능)



로지스틱 회귀분석

Heart Disease Prediction Binary Classification

- num:
 - 0 = no heart disease
 - 1 = mild heart disease
 - 2 = moderate heart disease
 - 3 = severe heart disease
 - 4 = critical heart disease

Dataset Description:

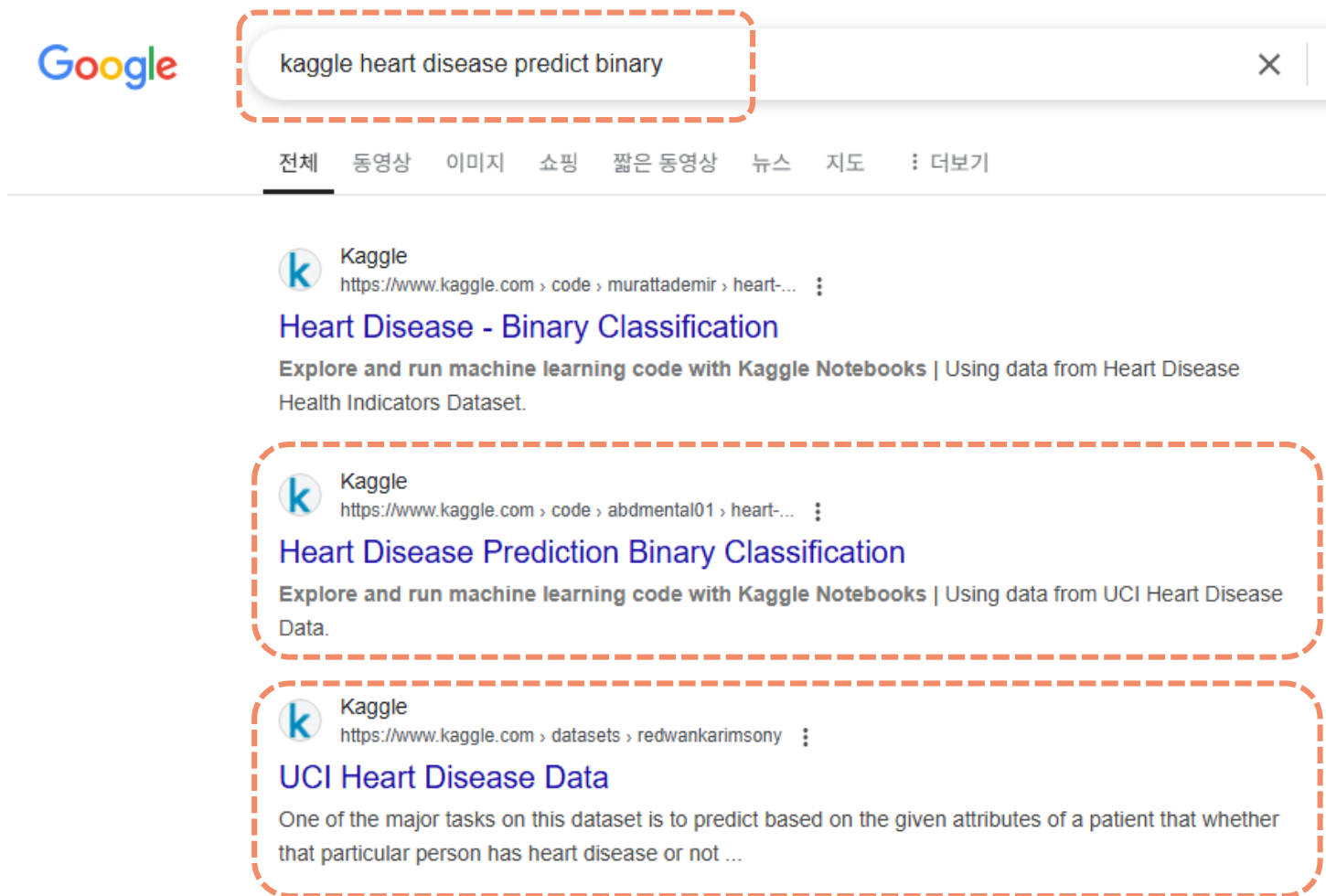
Variable	Description
age	Age of the patient in years
sex	Gender of the patient (0 = male, 1 = female)
cp	Chest pain type: 0: Typical angina, 1: Atypical angina, 2: Non-anginal pain, 3: Asymptomatic
trestbps	Resting blood pressure in mm Hg
chol	Serum cholesterol in mg/dl
fbs	Fasting blood sugar level, categorized as above 120 mg/dl (1 = true, 0 = false)
restecg	Resting electrocardiographic results: 0: Normal, 1: Having ST-T wave abnormality, 2: Showing probable or definite left ventricular hypertrophy
thalach	Maximum heart rate achieved during a stress test
exang	Exercise-induced angina (1 = yes, 0 = no)
oldpeak	ST depression induced by exercise relative to rest
slope	Slope of the peak exercise ST segment: 0: Upsloping, 1: Flat, 2: Downsloping
ca	Number of major vessels (0-4) colored by fluoroscopy
thal	Thalium stress test result: 0: Normal, 1: Fixed defect, 2: Reversible defect, 3: Not described
target	Heart disease status (0 = no disease, 1 = presence of disease)

<https://www.kaggle.com/code/abdmental01/heart-disease-prediction-binary-classification/notebook>

04 분류 모델을 이용한 질병 예측

■ 로지스틱 회귀분석

▪ 데이터 내려받기



The screenshot shows a Google search interface. The search bar contains the text "kaggle heart disease predict binary". Below the search bar, there are tabs for "전체", "동영상", "이미지", "쇼핑", "짧은 동영상", "뉴스", "지도", and "더보기". The search results are displayed below the tabs. The first result is from Kaggle, titled "Heart Disease - Binary Classification", with a description "Explore and run machine learning code with Kaggle Notebooks | Using data from Heart Disease Health Indicators Dataset." The second result is also from Kaggle, titled "Heart Disease Prediction Binary Classification", with a description "Explore and run machine learning code with Kaggle Notebooks | Using data from UCI Heart Disease Data." The third result is from Kaggle, titled "UCI Heart Disease Data", with a description "One of the major tasks on this dataset is to predict based on the given attributes of a patient that whether that particular person has heart disease or not ...". The second and third results are highlighted with orange dashed boxes.

Google

kaggle heart disease predict binary

전체 동영상 이미지 쇼핑 짧은 동영상 뉴스 지도 더보기

Kaggle
https://www.kaggle.com › code › murattademir › heart-...
Heart Disease - Binary Classification
Explore and run machine learning code with Kaggle Notebooks | Using data from Heart Disease Health Indicators Dataset.


Kaggle
https://www.kaggle.com › code › abdmental01 › heart-...
Heart Disease Prediction Binary Classification
Explore and run machine learning code with Kaggle Notebooks | Using data from UCI Heart Disease Data.

Kaggle
https://www.kaggle.com › datasets › redwankarimsony
UCI Heart Disease Data
One of the major tasks on this dataset is to predict based on the given attributes of a patient that whether that particular person has heart disease or not ...

04 분류 모델을 이용한 질병 예측

로지스틱 회귀분석

데이터 내려받기

 SHEIKH MUHAMMAD ABDULLAH · 9MO AGO · 6,781 VIEWS

184


Copy & Edit

166

⋮

Heart Disease Prediction Binary Classification

Notebook Input Output Logs Comments (18)



Heart Disease Prediction Project

Version 3 of 3

Runtime

7s

Input

DATASETS

heart-disease-data

UCI Heart Disease Data

Python

Table of Contents

About Data

Meta-Data (About Dataset):

Context:

ABOUT DATA

■ 로지스틱 회귀분석

▪ 데이터 내려받기

Input Data



UCI Heart Disease Data

Heart Disease Data Set from UCI data repository

Last Updated: 4 years ago (Version 6)

About this Dataset

Context

This is a multivariate type of dataset which means providing or involving a variety of separate mathematical or statistical variables, multivariate numerical data analysis. It is composed of 14 attributes which are age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, oldpeak — ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels and Thalassemia. This database includes 76 attributes, but all published studies relate to the use of a subset of 14 of them. The Cleveland database is the only one used by ML researchers to date. One of the major tasks on this dataset is to predict based on the given attributes of a patient that whether that particular person has heart disease or not and other is the experimental task to diagnose and find out various insights from this dataset which could help in understanding the problem more.

Content

04 분류 모델을 이용한 질병 예측

■ 로지스틱 회귀분석

▪ 데이터 내려받기

3. `origin` (place of study)
4. `sex` (Male/Female)
5. `cp` chest pain type ([typical angina, atypical angina, non-anginal, asymptomatic])

▼ View more

heart_disease_uci.csv (79.35 kB)

Detail Compact Column

10 of 16 columns ▼

[Download](#)

[Suggest Edits](#)

About this file

The original dataset had some numerical values which are categorical values that i changed with the original value.

■ 로지스틱 회귀분석

▪ 데이터 불러오기(csv)

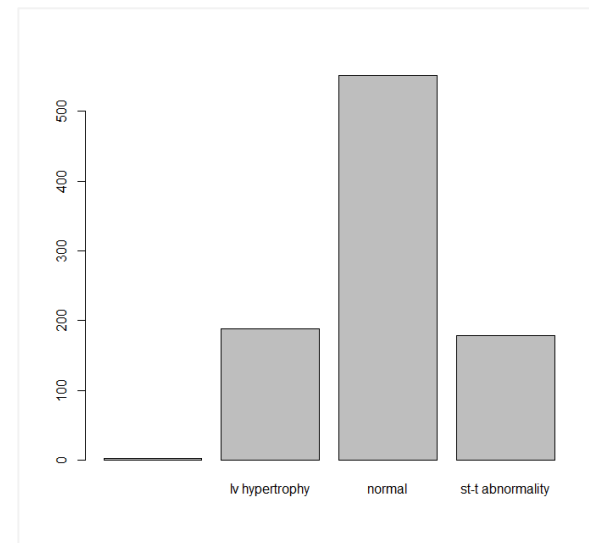
```
setwd("file directory") ## R 디폴트는 내 문서
data<-read.csv("heart_disease_uci.csv")
str(data)
```

```
> data<-read.csv("heart_disease_uci.csv")
> str(data)
'data.frame':   920 obs. of  16 variables:
 $ id       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ age      : int  63 67 67 37 41 56 62 57 63 53 ...
 $ sex      : chr  "Male" "Male" "Male" "Male" ...
 $ dataset  : chr  "Cleveland" "Cleveland" "Cleveland" "Cleveland" ...
 $ cp       : chr  "typical angina" "asymptomatic" "asymptomatic" "non-anginal" ...
 $ trestbps : int  145 160 120 130 130 120 140 120 130 140 ...
 $ chol     : int  233 286 229 250 204 236 268 354 254 203 ...
 $ fbs      : logi  TRUE FALSE FALSE FALSE FALSE FALSE ...
 $ restecg  : chr  "lv hypertrophy" "lv hypertrophy" "lv hypertrophy" "normal" ...
 $ thalch   : int  150 108 129 187 172 178 160 163 147 155 ...
 $ exang    : logi  FALSE TRUE TRUE FALSE FALSE FALSE ...
 $ oldpeak  : num  2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...
 $ slope    : chr  "downsloping" "flat" "flat" "downsloping" ...
 $ ca       : int  0 3 2 0 0 0 2 0 1 0 ...
 $ thal     : chr  "fixed defect" "normal" "reversable defect" "normal" ...
 $ num      : int  0 2 1 0 0 0 3 0 2 1 ...
```

■ 로지스틱 회귀분석

- 데이터 탐색
- `table()`: 빈도 확인(numeric, character 모두 가능)
- `hist()`: numeric 데이터에 유용
- `barplot()`: character 데이터에 유용

ex) `barplot(table(data$restecg))`



- (교재나 사람에 따라 히스토그램과 막대그래프를 다른 종류의 그래프로 간주하는 경우도 있음)

04 분류 모델을 이용한 질병 예측

■ 패키지 설치 및 로드

▪ ex) ggplot2 패키지 설치 후 부착

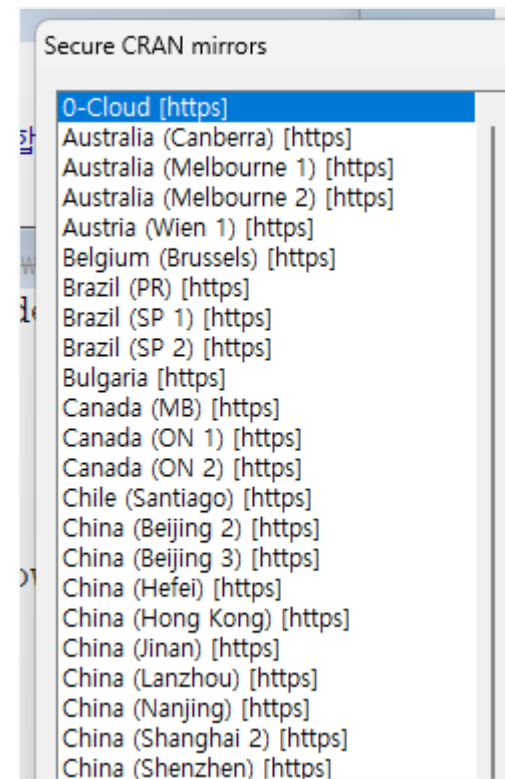
```
install.packages("ggplot2")  
library(ggplot2) ## 또는 require(ggplot2)
```

```
> install.packages("ggplot2")  
'C:/Users/AIM/AppData/Local/R/win-library/4.4'의 위치에 패키지(들)을 설치합니다.  
(왜냐하면 'lib'가 지정되지 않았기 때문입니다)  
--- 현재 세션에서 사용할 CRAN 미러를 선택해 주세요 ---  
URL 'https://cloud.r-project.org/bin/windows/contrib/4.4/ggplot2_3.5.1.zip'을 시도합니다  
Content type 'application/zip' length 5021782 bytes (4.8 MB)  
downloaded 4.8 MB
```

패키지 'ggplot2'를 성공적으로 압축해제하였고 MD5 sums 이 확인되었습니다

다운로드된 바이너리 패키지들은 다음의 위치에 있습니다

C:\Users\AIM\AppData\Local\Temp\RtmpcdIkKW\downloaded_packages



#chooseCRANmirror()로
언제든 변경 가능

로지스틱 회귀분석

데이터 전처리

1. 결측치 처리

*로지스틱 회귀분석은 결측치가 있을 경우 분석이 불가능 (샘플에서 자동 제외)

`is.na(data)` #각 샘플의 변수 별 결측 여부 확인

```
      id   age  sex dataset    cp trestbps  chol   fbs restecg thalch exang
[1,] FALSE FALSE FALSE  FALSE FALSE      FALSE FALSE FALSE  FALSE  FALSE FALSE
[2,] FALSE FALSE FALSE  FALSE FALSE      FALSE FALSE FALSE  FALSE  FALSE FALSE
[3,] FALSE FALSE FALSE  FALSE FALSE      FALSE FALSE FALSE  FALSE  FALSE FALSE
[597,] FALSE FALSE FALSE  FALSE FALSE      FALSE FALSE FALSE  FALSE  FALSE FALSE
[598,] FALSE FALSE FALSE  FALSE FALSE      FALSE FALSE  TRUE  FALSE  FALSE FALSE
[599,] FALSE FALSE FALSE  FALSE FALSE      FALSE FALSE  TRUE  FALSE  FALSE FALSE
[600,] FALSE FALSE FALSE  FALSE FALSE      TRUE  FALSE  TRUE  FALSE  FALSE FALSE
[601,] FALSE FALSE FALSE  FALSE FALSE      FALSE FALSE  TRUE  FALSE  FALSE FALSE
[602,] FALSE FALSE FALSE  FALSE FALSE      FALSE FALSE  TRUE  FALSE  FALSE FALSE
[603,] FALSE FALSE FALSE  FALSE FALSE      FALSE FALSE  FALSE  FALSE  FALSE FALSE
```

■ 로지스틱 회귀분석

▪ 데이터 전처리

1. 결측치 처리

```
> dim(data)           # 결측치 제거 전 샘플 920개 변수 16개  
[1] 920 16  
> data2<-na.omit(data) # 결측치 있는 샘플 제거  
> dim(data2)  
[1] 303 16           # 결측치 제거 후 샘플 303개
```

****만약 결측치가 NA외의 “”(blank, 공란) 등의 다른 값으로 설정되어 있다면?**

■ 로지스틱 회귀분석

- 데이터 전처리

1. 결측치 처리(dplyr 패키지 사용)

```
data2<-filter(data2, if_all(everything(), ~!is.na(.) & .!=""))
```

-----	-----	-----	-----
# 특정 조건을 만족하는 케이스만 추출	#모든 케이스	#NA가 아님	# "" (Blank)가 아님

■ 로지스틱 회귀분석

▪ 데이터 전처리

2. 이상치 처리

- EDA(탐색적 자료 분석) 도중 이상치를 발견하는 경우

(기준상한, 하한을 벗어나는 경우 이상치로 판정, $\pm k\sigma$ 가 주로 이용됨)

우리는 이 데이터가 (1)잘못 나온 수치인지, (2)잘못 입력한 수치인지,

(3)실제로 나온 수치인지 알 수 없음

데이터 관리자에게 연락해 확인하는 과정이 가장 적절

이상치 처리 방법

- 이상치 삭제

- 다른 값 대체

- 그대로 두기(*)

**몇몇 가이드라인은 제시되어 있지만, 결국 연구자가 판단하여 결정해야 함

로지스틱 회귀분석

- 데이터 전처리

3. 변수 변환

결과 변수 num은 횡수이므로

ifelse 함수를 통해 유, 무 형태로 변환

```
> data2$num
[1] 0 2 1 0 0 0 3 0 2 1 0 0 2 0 0 0 1 0 0 0 0 0 1 3 4 0 0 0 0 3 0 2 1 0 0 0 3
[38] 1 3 0 4 0 0 0 1 4 0 4 0 0 0 0 2 0 1 1 1 1 0 0 2 0 1 0 2 2 1 0 2 1 0 3 1 1
[75] 1 0 1 0 0 3 0 0 0 3 0 0 0 0 0 0 3 0 0 0 1 2 3 0 0 0 0 0 0 3 0 2 1 2 3 1 1
[112] 0 2 2 0 0 0 3 2 3 4 0 3 1 0 3 3 0 0 0 0 0 0 0 0 0 4 3 1 0 0 1 0 1 0 1 4 0 0
[149] 0 0 0 0 4 3 1 1 1 2 0 0 4 0 0 0 0 0 1 0 3 0 1 0 4 1 0 1 0 0 3 2 0 0 1 0 0
[186] 2 1 2 0 3 2 0 3 0 0 0 1 0 0 0 0 0 3 3 3 0 1 0 4 0 3 1 0 0 0 0 0 0 0 0 3 1
[223] 0 0 0 3 2 0 2 1 0 0 3 2 1 0 0 0 0 0 2 0 2 2 1 3 0 0 1 0 0 0 0 0 0 0 1 0 3
[260] 0 0 4 2 2 1 0 1 0 2 0 1 0 0 0 1 0 2 0 3 0 2 4 2 0 0 1 0 2 2 1 0 3 1 1 2 3
[297] 1 1 1
```

data2\$target<-ifelse(data2\$num>0,1,0)

```
> data2$target
[1] 0 1 1 0 0 0 1 0 1 1 0 0 1 0 0 0 1 0 0 0 0 0 1 1 1 0 0 0 0 1 0 1 1 0 0 0 1
[38] 1 1 0 1 0 0 0 1 1 0 1 0 0 0 0 1 0 1 1 1 1 0 0 1 0 1 0 1 1 1 0 1 1 0 1 1 1
[75] 1 0 1 0 0 1 0 0 0 1 0 0 0 0 0 0 1 0 0 0 1 1 1 0 0 0 0 0 0 1 0 1 1 1 1 1 1
[112] 0 1 1 0 0 0 1 1 1 1 0 1 1 0 1 1 0 0 0 0 0 0 0 0 0 1 1 1 0 0 1 0 1 0 1 1 0 0
[149] 0 0 0 0 1 1 1 1 1 1 0 0 1 0 0 0 0 0 1 0 1 0 1 0 1 1 0 1 0 0 1 1 0 0 1 0 0
[186] 1 1 1 0 1 1 0 1 0 0 0 1 0 0 0 0 0 1 1 1 0 1 0 1 0 1 1 0 0 0 0 0 0 0 0 1 1
[223] 0 0 0 1 1 0 1 1 0 0 1 1 1 0 0 0 0 0 1 0 1 1 1 1 0 0 1 0 0 0 0 0 0 0 1 0 1
[260] 0 0 1 1 1 1 0 1 0 1 0 1 0 0 0 1 0 1 0 1 0 1 1 1 0 0 1 0 1 1 1 0 1 1 1 1 1
[297] 1 1 1
```


■ 로지스틱 회귀분석

■ 모형 적합

```
model<-glm(target ~ age+sex+dataset+cp+trestbps+chol+  
            fbs+restecg+thalch+exang+oldpeak+slope+  
            ca+thal, family=binomial, data=data2)  
summary(model) -----  
# 일반화 선형 모형에서 로지스틱 회귀분석을 시행하는 옵션
```

■ 또는

```
data2sub<-subset(data2, select=c(-id, -num)) # 데이터셋에서 해당 변수 제거  
colnames(data2sub)  
model2<-glm(target ~ ., family=binomial, data=data2sub)  
summary(model2) -----  
# 나머지 모든 변수를 모형에 포함하겠다는 뜻
```

로지스틱 회귀분석

결과 해석

R의 glm() 함수는

자동으로 범주형 변수를

one-hot encoding하여 인식

Reference value는

따로 level을 설정하지 않으면

알파벳 오름차순으로 설정됨

```
Call:
glm(formula = target ~ ., family = binomial, data = data2sub)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.95492    2.84635  -1.04   0.29920
age           -0.01376    0.02475  -0.56   0.57808
# ref : female sexMale      1.54601    0.53000   2.92   0.00353 **
# ref : Cleveland datasetHungary 11.89258 1455.39778   0.01   0.99348
datasetVA Long Beach 13.69774 1455.39773   0.01   0.99249
cpatypical angina  -0.84691    0.56046  -1.51   0.13076
# ref : asymptomatic cpnon-anginal -1.84052    0.50058  -3.68   0.00024 ***
cptypical angina  -2.08648    0.66655  -3.13   0.00175 **
trestbps      0.02436    0.01127   2.16   0.03062 *
chol          0.00445    0.00399   1.11   0.26526
fbsTRUE      -0.59625    0.60785  -0.98   0.32664
# ref : lv hypertrophy restecgnormal -0.47389    0.38352  -1.24   0.21659
restecgst-t abnormality 0.33631    2.43742   0.14   0.89026
thalch       -0.01772    0.01111  -1.60   0.11065
# ref : FALSE exangTRUE    0.70946    0.44002   1.61   0.10689
oldpeak      0.35787    0.23007   1.56   0.11983
# ref : downsloping slopeflat    0.63014    0.84829   0.74   0.45758
slopeupsloping -0.52515    0.91966  -0.57   0.56799
ca           1.31151    0.27928   4.70 0.0000027 ***
# ref : fixed defect thalnormal    0.01097    0.79021   0.01   0.98892
thalreversible defect 1.40369    0.77614   1.81   0.07052 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 413.03  on 298  degrees of freedom
Residual deviance: 191.64  on 278  degrees of freedom
AIC: 233.6
```

로지스틱 회귀분석

결과 해석

로지스틱 회귀 모델에서 estimate(추정값)의 계수는 로그 오즈(log-odds)를 의미함

범주형 변수의 경우

ex) Sex = Male의 경우 추정값이 $\exp(1.54)$

$\exp(1.54) = 4.665$

남성이 ref(Female)에 비해 위험 4.665배 높음

```
Call:
glm(formula = target ~ ., family = binomial, data = data2sub)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.95492    2.84635  -1.04  0.29920
age           -0.01376    0.02475  -0.56  0.57808
sexMale        1.54601    0.53000   2.92  0.00353 **
datasetHungary 11.89258 1455.39778   0.01  0.99348
datasetVA Long Beach 13.69774 1455.39773   0.01  0.99249
cpatypical angina -0.84691    0.56046  -1.51  0.13076
cpnon-anginal   -1.84052    0.50058  -3.68  0.00024 ***
cptypical angina -2.08648    0.66655  -3.13  0.00175 **
trestbps        0.02436    0.01127   2.16  0.03062 *
chol           0.00445    0.00399   1.11  0.26526
fbsTRUE        -0.59625    0.60785  -0.98  0.32664
restecgnormal  -0.47389    0.38352  -1.24  0.21659
restecgst-t abnormality 0.33631    2.43742   0.14  0.89026
thalch         -0.01772    0.01111  -1.60  0.11065
exangTRUE       0.70946    0.44002   1.61  0.10689
oldpeak        0.35787    0.23007   1.56  0.11983
slopeflat      0.63014    0.84829   0.74  0.45758
slopeupsloping -0.52515    0.91966  -0.57  0.56799
ca             1.31151    0.27928   4.70 0.0000027 ***
thalnormal     0.01097    0.79021   0.01  0.98892
thalreversible defect 1.40369    0.77614   1.81  0.07052 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 413.03  on 298  degrees of freedom
Residual deviance: 191.64  on 278  degrees of freedom
AIC: 233.6
```

로지스틱 회귀분석

결과 해석

연속형 변수의 경우

ex) trestbps의 경우 1단위 상승할 때 마다

위험률 $\exp(0.02436) = 1.025$ 배 증가

```
Call:
glm(formula = target ~ ., family = binomial, data = data2sub)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.95492    2.84635   -1.04  0.29920
age           -0.01376    0.02475   -0.56  0.57808
sexMale        1.54601    0.53000    2.92  0.00353 **
datasetHungary 11.89258 1455.39778    0.01  0.99348
datasetVA Long Beach 13.69774 1455.39773    0.01  0.99249
cpatypical angina -0.84691    0.56046   -1.51  0.13076
cpnon-anginal  -1.84052    0.50058   -3.68  0.00024 ***
cptypical angina -2.08648    0.66655   -3.13  0.00175 **
trestbps        0.02436    0.01127    2.16  0.03062 *
chol           0.00445    0.00399    1.11  0.26526
fbsTRUE        -0.59625    0.60785   -0.98  0.32664
restecgnormal  -0.47389    0.38352   -1.24  0.21659
restecgst-t abnormality 0.33631    2.43742    0.14  0.89026
thalch         -0.01772    0.01111   -1.60  0.11065
exangTRUE       0.70946    0.44002    1.61  0.10689
oldpeak        0.35787    0.23007    1.56  0.11983
slopeflat       0.63014    0.84829    0.74  0.45758
slopeupsloping  -0.52515    0.91966   -0.57  0.56799
ca              1.31151    0.27928    4.70 0.0000027 ***
thalnormal      0.01097    0.79021    0.01  0.98892
thalreversible defect 1.40369    0.77614    1.81  0.07052 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 413.03  on 298  degrees of freedom
Residual deviance: 191.64  on 278  degrees of freedom
AIC: 233.6
```

로지스틱 회귀분석

오즈비

두 그룹 간 특정 사건이 발생할 가능성을
비교하는 지표

한 그룹에서 사건이 발생할 오즈(odds)가
다른 그룹에 비해 몇 배 높은가?

	event 0	event X
Group A	a	b
Group B	c	d

각 그룹의 오즈는

$$\text{Group A} = \frac{a}{b}$$

$$\text{Group B} = \frac{c}{d}$$

$$\text{Odds Ratio} = \frac{a/b}{c/d} = \frac{a \times d}{b \times c}$$

로지스틱 회귀분석

오즈비

- cross table을 통해 직접 계산

```
table(data2$sex, data2$target)
```

	0	1
Female	72	25
Male	91	115

남성의 사건 발생 오즈
 $91/115 = 1.2637$

여성의 사건 발생 오즈

$25/72 = 0.3472$

오즈비는

$$OR = \frac{Odds_{male}}{Odds_{female}} = \frac{1.2637}{0.3472} = 3.64$$

⇒ 남성의 사건발생률이 여성에 비해 3.64배이다.

로지스틱 회귀분석

로지스틱 회귀분석 결과와 비교

- 교차표를 통해 구한 오즈비: 3.64
- 로지스틱 회귀 분석을 통해 구한 오즈비: $\exp(1.546) = 4.665$
- 교차표를 통해 구한 오즈비를 Crude Odds Ratio,

로지스틱 회귀 분석을 통해 구한 오즈비를 Adjusted Odds Ratio로 표기

- 보통 단순 로짓 회귀분석(독립변수 종속변수 1 on 1)을 통해 구한 오즈비를 Crude Odds Ratio로 간주하지만, 교란변수 또는 다층 구조 등이 있을 경우 교차표를 통해 구한 오즈비와 다를 수 있기 때문에 주의해야 함

	estimate	
(Intercept)	-2.95492	
age	-0.01376	
sexMale	1.54601	
datasetHungary	11.89258	1
datasetVA Long Beach	13.69774	1

Cummings, P. (2009). Methods for Estimating Adjusted Risk Ratios. *The Stata Journal: Promoting Communications on Statistics and Stata*, 9(2), 175–196.

<https://doi.org/10.1177/1536867X0900900201>

■ 로지스틱 회귀분석

- 실습)

Cross table로 직접 계산한 오즈비,

독립변수와 종속변수의 1:1 로지스틱 회귀분석 시행 결과의 오즈비,

다중 로지스틱 회귀분석 시행 결과의 오즈비

셋을 비교해보자. 어떠한 차이가 있을까?

로지스틱 회귀분석

- 결과 해석
- Null deviance vs. Residual deviance

독립변수를 하나도 사용하지 않은 경우와
현재 모델의 이탈도 비교

모형의 유의성 검정에 사용됨

$$H_0 : \beta = 0 \text{ vs. } H_1 = \text{not } H_0$$

$$LR = \text{Null deviance} - \text{Residual deviance}$$

$$\sim \chi^2_{df_{null}-df_{residual}}$$

```
Call:
glm(formula = target ~ ., family = binomial, data = data2sub)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.95492    2.84635   -1.04  0.29920
age           -0.01376    0.02475   -0.56  0.57808
sexMale        1.54601    0.53000    2.92  0.00353 **
datasetHungary 11.89258 1455.39778    0.01  0.99348
datasetVA Long Beach 13.69774 1455.39773    0.01  0.99249
cpatypical angina -0.84691    0.56046   -1.51  0.13076
cpnon-anginal  -1.84052    0.50058   -3.68  0.00024 ***
cptypical angina -2.08648    0.66655   -3.13  0.00175 **
trestbps        0.02436    0.01127    2.16  0.03062 *
chol           0.00445    0.00399    1.11  0.26526
fbsTRUE        -0.59625    0.60785   -0.98  0.32664
restecgnormal  -0.47389    0.38352   -1.24  0.21659
restecgst-t abnormality 0.33631    2.43742    0.14  0.89026
thalch         -0.01772    0.01111   -1.60  0.11065
exangTRUE       0.70946    0.44002    1.61  0.10689
oldpeak        0.35787    0.23007    1.56  0.11983
slopeflat      0.63014    0.84829    0.74  0.45758
slopeupsloping -0.52515    0.91966   -0.57  0.56799
ca             1.31151    0.27928    4.70 0.0000027 ***
thalnormal      0.01097    0.79021    0.01  0.98892
thalreversible defect 1.40369    0.77614    1.81  0.07052 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 413.03  on 298  degrees of freedom
Residual deviance: 191.64  on 278  degrees of freedom
AIC: 233.6
```

로지스틱 회귀분석

- 변수 선택
- 단계선택법

BIC를 기준으로 할 경우

$$k = \log(\text{nrow}(\text{data2sub}))$$

`summary(step(model2, direction="both", k=2))`

- 일부 변수가 삭제되고 AIC가 감소
유익한 변수도 일부 추가

**** BIC는 더 큰 패널티로
더 단순한 모델을 선택하는 경향이 있음**

```
Call:
glm(formula = target ~ sex + cp + trestbps + thalch + exang +
     oldpeak + slope + ca + thal, family = binomial, data = data2sub)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -3.2734     2.2478   -1.46   0.1453
sexMale         1.4385     0.4942    2.91   0.0036 **
cpatypical angina -0.8820     0.5519   -1.60   0.1100
cpnon-anginal   -1.9773     0.4876   -4.06 0.0000500 ***
cptypical angina -2.1819     0.6573   -3.32   0.0009 ***
trestbps         0.0229     0.0103    2.22   0.0265 *
thalch          -0.0152     0.0101   -1.50   0.1327
exangTRUE        0.6749     0.4328    1.56   0.1189
oldpeak          0.3937     0.2227    1.77   0.0771 .
slopeflat        0.7556     0.8303    0.91   0.3628
slopeupsloping  -0.4490     0.8964   -0.50   0.6165
ca               1.2370     0.2571    4.81 0.0000015 ***
thalnormal       0.2475     0.7618    0.32   0.7453
thalreversible defect 1.6152     0.7475    2.16   0.0307 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 413.03  on 298  degrees of freedom
Residual deviance: 196.16  on 285  degrees of freedom
AIC: 224.2
```

■ 로지스틱 회귀분석

▪ 변수 중요도 분석

모델의 예측 결과에 각 독립변수가 얼마나
영향을 미쳤는지 평가하는 방법 중 하나
특정 변수가 종속변수에 미치는 상대적 영향력을
정량적으로 측정

```
varImp(step(model2, direction="both"))
```

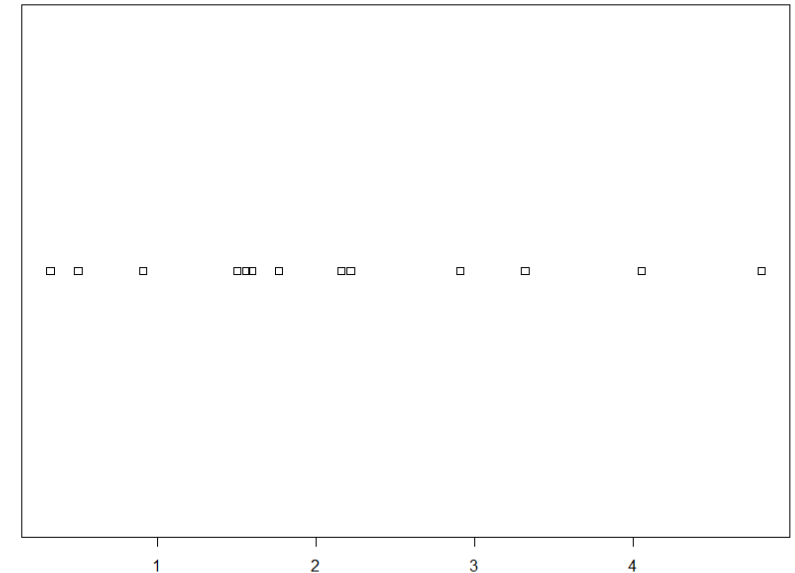
	Overall
sexMale	2.9106
cpatypical angina	1.5982
cpnon-anginal	4.0554
cptypical angina	3.3197
trestbps	2.2192
thalch	1.5035
exangTRUE	1.5594
oldpeak	1.7680
slopeflat	0.9101
slopeupsloping	0.5008
ca	4.8110
thalnormal	0.3249
thalreversible defect	2.1609

로지스틱 회귀분석

변수 중요도 그래프

plot() 명령어를 사용시
varImp() 결과의
Overall값만 반환되며
오른쪽과 같은 결과가
나타남

	Overall
sexMale	2.9106
cpatypical angina	1.5982
cpnon-anginal	4.0554
cptypical angina	3.3197
trestbps	2.2192
thalch	1.5035
exangTRUE	1.5594
oldpeak	1.7680
slopeflat	0.9101
slopeupsloping	0.5008
ca	4.8110
thalnormal	0.3249
thalreversable defect	2.1609

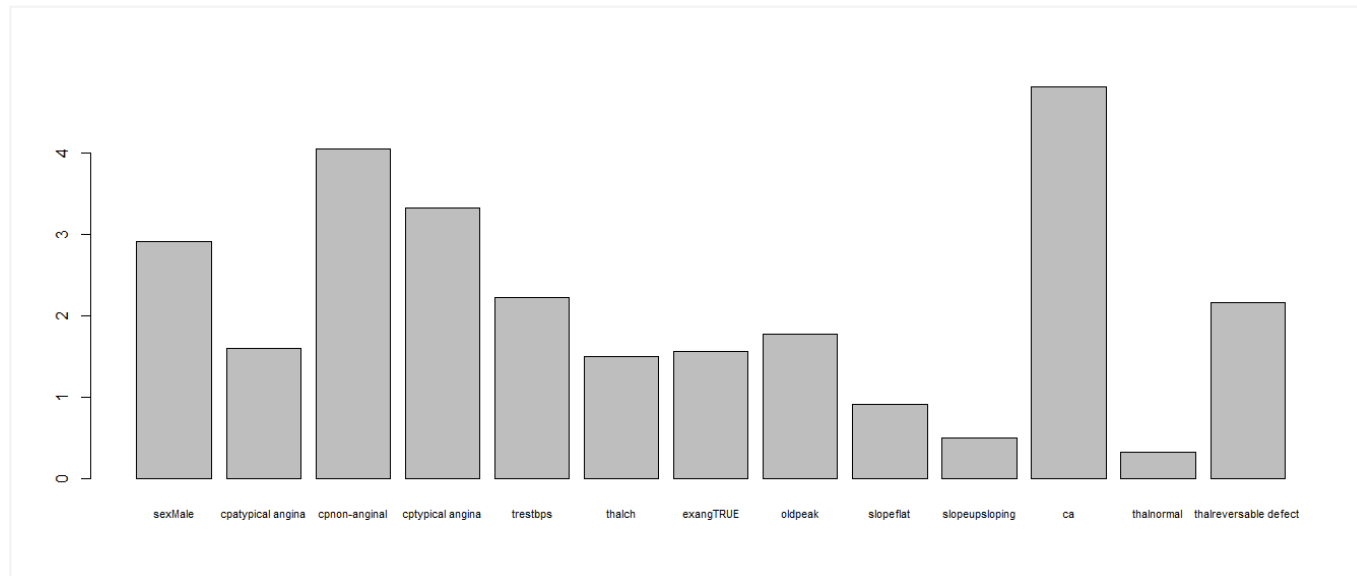


로지스틱 회귀분석

- 변수 중요도 그래프
- 막대그래프를 통해 나타낼 수 있음

```
barplot(names.arg=rownames(vi),  
        vi$Overall, cex.names=0.7)
```

- horiz 옵션을 통해 축 변경 가능(horiz=T)



■ 로지스틱 회귀분석

- 변수 중요도 그래프
- ggplot2 패키지 이용

```
var_imp_df <- data.frame(Variable = rownames(vi),  
                          Importance = vi$Overall)
```

- “Variable”, “Importance”를 각각
변수명으로 갖는 데이터 프레임 형태로 변환

```
> var_imp_df
```

	Variable	Importance
1	sexMale	2.9106
2	cpatypical angina	1.5982
3	cpnon-anginal	4.0554
4	cptypical angina	3.3197
5	trestbps	2.2192
6	thalch	1.5035
7	exangTRUE	1.5594
8	oldpeak	1.7680
9	slopeflat	0.9101
10	slopeupsloping	0.5008
11	ca	4.8110
12	thalnormal	0.3249
13	thalreversible defect	2.1609

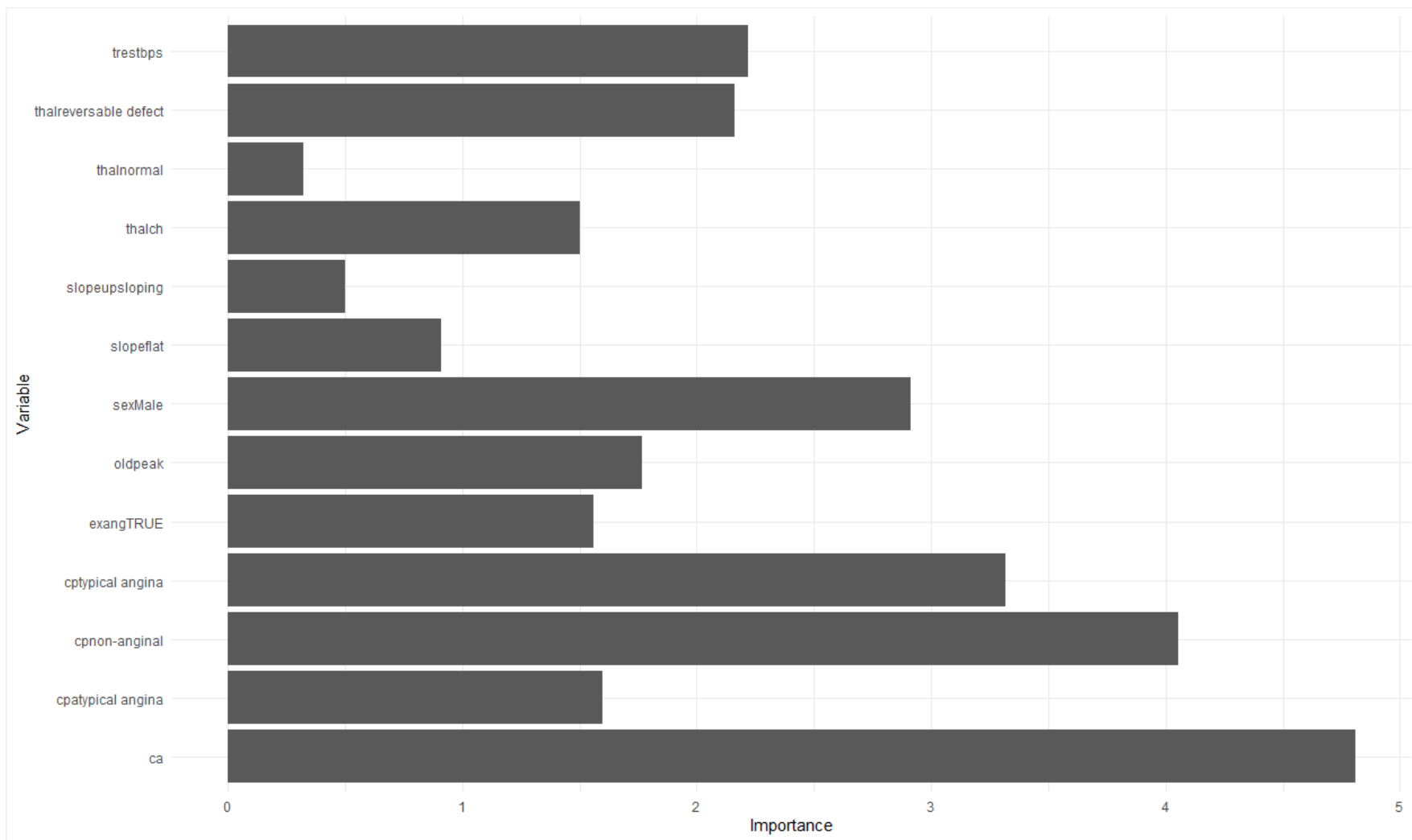
- 로지스틱 회귀분석
 - 변수 중요도 그래프

```
ggplot(var_imp_df, aes(x = Variable, y = Importance)) +  
  coord_flip() +  
  geom_bar(stat = "identity") +  
  theme_minimal()
```

Annotations:

- #그래프 축 변경 (points to `coord_flip()`)
- #가로 및 세로 설정 (points to `aes(x = Variable, y = Importance)`)
- #막대 그래프 (points to `geom_bar(stat = "identity")`)
- #value를 output으로 사용하는 옵션 (points to `stat = "identity"`)

04 분류 모델을 이용한 질병 예측



■ 로지스틱 회귀분석

- Train Set & Test Set
 - 모델을 훈련(Train)하고, 훈련된 모델을 검증(Test) 데이터에 사용하여 성능을 평가하여 모델의 일반화 성능 및 신뢰성을 확보하는 것이 중요
 - Train/Test Split을 수행하는 이유
 - 1. 과적합 방지 – 훈련 데이터만 너무 학습하면 새로운 데이터에 대해 성능이 떨어짐
 - 2. 객관적인 평가 – 새로운 데이터에서도 잘 작동해야 모델의 의미가 있음
- ⇒ Real Data 적용 가능성 확보 – 훈련 데이터만 평가하면 실제 배포 시 성능 하락 가능성 큼

로지스틱 회귀분석

Train Set & Test Set

#재현성 확보의 중요성

- 시드를 설정하면 동일한 코드 실행 시 항상 같은 결과가 나옴
- 동일한 실험을 다시 수행할 수 있어야 신뢰도가 높아짐
- 연구, 논문, 산업에 있어서 일관된 성능 보장 확보

```
set.seed(42) #재현성 확보를 위한 난수 생성 시작점 고정
library(caret) #createDataPartition 함수가 포함된 패키지
train_index<-createDataPartition(data2sub$target, p=0.7, list=FALSE)
train_data<-data2sub[train_index, ] #훈련 데이터셋의 index에 따라 훈련 데이터 할당
test_data<-data2sub[-train_index, ] #훈련 데이터셋의 index가 아닌 샘플들은 테스트 데이터 할당
```

#랜덤하게 70% 훈련 세트를 뽑는 명령어
반환값은 샘플 데이터셋의 index

```
dim(data2sub)
dim(train_data)
dim(test_data)
```

```
> dim(data2sub)    > dim(train_data)    > dim(test_data)
[1] 299  15        [1] 210  15        [1] 89  15
```

#샘플 개수는 정수이므로, 항상 정확한 비율로 나누어 떨어지지 않는음

로지스틱 회귀분석

- 훈련 데이터셋 로지스틱 회귀 모형 적합

```
train_model<-glm(target ~.,  
                  family="binomial",  
                  data=train_data)  
summary(step(train_model))
```

- 샘플 데이터가 달라졌기 때문에
적합 결과도 달라짐

```
Call:  
glm(formula = target ~ sex + cp + trestbps + thalch + ca + thal,  
     family = "binomial", data = train_data)  
  
Coefficients:  
                Estimate Std. Error z value Pr(>|z|)  
(Intercept)      0.93630    2.23448   0.419 0.675198  
sexMale           1.34335    0.53550   2.509 0.012122 *  
cpatypical angina -2.07374    0.72658  -2.854 0.004315 **  
cpnon-anginal     -1.91600    0.52652  -3.639 0.000274 ***  
cptypical angina  -2.05929    0.67907  -3.033 0.002425 **  
trestbps           0.02224    0.01176   1.891 0.058577 .  
thalch            -0.03188    0.01123  -2.840 0.004516 **  
ca                1.13329    0.27819   4.074 4.63e-05 ***  
thalnormal        -0.50179    0.82312  -0.610 0.542112  
thalreversible defect 1.20754    0.80665   1.497 0.134399  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
Null deviance: 290.95  on 209  degrees of freedom  
Residual deviance: 142.05  on 200  degrees of freedom  
AIC: 162.05  
  
Number of Fisher Scoring iterations: 6
```

로지스틱 회귀분석

로지스틱 회귀분석에서의 다중공선성

GVIF(Generalized VIF)

- 범주형 변수를 포함한 모델에서
사용 가능하도록 확장된 VIF

- 독립변수의 자유도가 2보다 클 경우

보정된 값($GVIF^{\frac{1}{2 \times Df}}$)을 해석

```
vif(step(train_model))
```

	GVIF	Df	$GVIF^{(1/(2 \times Df))}$
sex	1.242790	1	1.114805
cp	1.433561	3	1.061865
trestbps	1.096560	1	1.047168
thalch	1.286023	1	1.134030
ca	1.132200	1	1.064049
thal	1.469023	2	1.100924

- GVIF가 2 또는 5보다 크면 다중공선성을
의심할 만하다고 하나 절대적인 기준은 없음

로지스틱 회귀분석

- 혼동행렬(Confusion Matrix)
- accuracy는 전체 샘플 중에서 올바르게 분류된 비율
각 샘플이 올바르게 분류될 확률을 가지는 이항분포를 따름

- $$\text{accuracy} = \frac{50+23}{50+8+8+23} = 0.8202 \text{ (약 82\%)}$$

- accuracy의 신뢰구간 : (0.7245, 0.8936)

- accuracy의 참 값이 신뢰구간 안에 있을 확률 : 95%

(정확하게는, 작업을 100번 진행하여 100개의 신뢰구간을 만들었을 때, 그 중 95개가 accuracy의 참 값을 포함한다는 의미)

Confusion Matrix and Statistics

```
Reference
Prediction 0 1
0 50 8
1 8 23
```

```
Accuracy : 0.8202
95% CI : (0.7245, 0.8936)
```

```
No Information Rate : 0.6517
P-Value [Acc > NIR] : 0.0003593
```

```
Kappa : 0.604
```

```
Mcnemar's Test P-Value : 1.0000000
```

```
Sensitivity : 0.8621
Specificity : 0.7419
Pos Pred Value : 0.8621
Neg Pred Value : 0.7419
Prevalence : 0.6517
Detection Rate : 0.5618
Detection Prevalence : 0.6517
Balanced Accuracy : 0.8020
```

```
'Positive' Class : 0
```

로지스틱 회귀분석

혼동행렬(Confusion Matrix)

No Information Rate

- 다수의 클래스를 무조건 예측했을 때 얻는 정확도

$$= \frac{58}{50+8+8+23} = 0.6518$$

Cohen's Kappa

- 모델이 랜덤 예측하는 경우를 보정한 정확도 측정 지표

$$\Rightarrow \kappa = \frac{P_0 - P_e}{1 - P_e}, P_e : \text{각 클래스의 실제비율과 예측비율 곱의 합}$$

$$P_e = (P_{actual0} \times P_{predicted0}) + (P_{actual1} \times P_{predicted1})$$

⇒ 0.6 이상이면 신뢰도가 높은 것으로 간주됨

Confusion Matrix and Statistics

	Reference	
Prediction	0	1
0	50	8
1	8	23

Accuracy : 0.8202

95% CI : (0.7245, 0.8936)

No Information Rate : 0.6517

P-Value [Acc > NIR] : 0.0003593

Kappa : 0.604

Mcnemar's Test P-Value : 1.0000000

Sensitivity : 0.8621

Specificity : 0.7419

Pos Pred Value : 0.8621

Neg Pred Value : 0.7419

Prevalence : 0.6517

Detection Rate : 0.5618

Detection Prevalence : 0.6517

Balanced Accuracy : 0.8020

'Positive' Class : 0

로지스틱 회귀분석

- 혼동행렬(Confusion Matrix)
- McNemar 검정

분류 모델의 예측 성능이 대칭적인가에 대한 검정

- 실제 0인데 1로 예측하는 사건을 b
실제 1인데 0으로 예측하는 사건을 c

$$Z = \frac{b-c}{\sqrt{b+c}} \sim N(0,1) \Rightarrow \chi^2 = \frac{(b-c)^2}{b+c}$$

- 본 예제에서는 분자가 0이므로 p-값은 1이 나옴
(아래 추가 내용에 대해서는 추후 설명할 예정)

Confusion Matrix and Statistics

```
Reference
Prediction 0 1
0 50 8
1 8 23
```

```
Accuracy : 0.8202
95% CI : (0.7245, 0.8936)
No Information Rate : 0.6517
P-Value [Acc > NIR] : 0.0003593
```

```
Kappa : 0.604
```

```
McNemar's Test P-Value : 1.0000000
```

```
Sensitivity : 0.8621
Specificity : 0.7419
Pos Pred Value : 0.8621
Neg Pred Value : 0.7419
Prevalence : 0.6517
Detection Rate : 0.5618
Detection Prevalence : 0.6517
Balanced Accuracy : 0.8020
```

```
'Positive' Class : 0
```


감사합니다

Q&A



충북대학교
CHUNGBUK NATIONAL UNIVERSITY