로지스틱 회귀분석(Logistic Regression Analysis)

http://ocw.ulsan.ac.kr/CourseLectures.aspx?CollCd=11161&DeptCd=11178&CourseNo=20101G0298501
9주차 로지스틱 회귀분석

종속변수와 독립변수 간의 관계를 나타내어 예측 모델을 생성한다는 점에서 선형 회귀분석 방법과 유사

로지스틱 회귀분석의 특징

- 분석 목적: 종속변수와 독립변수 간의 관계를 통해서 예측 모델 생성
- 회귀분석과 차이점: 종속변수는 반드시 범주형 변수(예, Yes/No, iris데이터의 species)
- 정규성: 정규분포 대신에 이항분포를 따른다.
- 로짓 변환: 종속변수의 출력범위를 0과 1로 조정하는 과정(예, 혈액형 A → [1, 0, 0, 0]
- 활용분야: 의료, 통신, 날씨 등 다양한 분야

실습 (날씨 관련 요인 변수로 비(rain) 유무 예측)

```
install.packages("ROCR")
library(car)
library(lmtest)
library(ROCR)
```

1단계: 데이터 가져오기

```
weather = read.csv("weather.csv", stringsAsFactors = F)
dim(weather)
head(weather)
str(weather)
```

2단계: 변수 선택과 더미 변수 생성

```
weather_df <- weather[ , c(-1, -6, -8, -14)]
str(weather_df)

weather_df$RainTomorrow[weather_df$RainTomorrow == 'Yes'] <- 1
weather_df$RainTomorrow[weather_df$RainTomorrow == 'No'] <- 0
```

```
weather_df$RainTomorrow <- as.numeric(weather_df$RainTomorrow)
head(weather_df)
```

X, Y변수 설정
Y변수를 대상으로 더미 변수를 생성하여 로지스틱 회귀분석 환경 설정

3단계: 학습데이터와 검정데이터 생성(7:3비율)

```
idx <- sample(1:nrow(weather_df), nrow(weather_df) * 0.7)
train <- weather_df[idx, ]
test <- weather_df[-idx, ]
```

4단계: 로지스틱 회귀모델 생성

```
weather_model <- glm(RainTomorrow ~ ., data = train, family = 'binomial', na.action=na.omit)
weather_model
summary(weather_model)
```

glm()함수
형식: glm(y~x, data, family)
Where
family = 'binomial' 속성: y변수가 이항형
https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm

로지스틱 회귀모델의 결과는 선형 회귀모델과 동일하게 x변수의 유의성 검정을 제공
하지만 F-검정 통계량과 모델의 설명력은 제공되지 않는다.

5단계: 로지스틱 회귀모델 예측치 생성

```
pred <- predict(weather_model, newdata = test, type = "response")
pred


result_pred <- ifelse(pred >= 0.5, 1, 0)
result_pred

table(result_pred)
```

predict() 함수

https://www.rdocumentation.org/packages/car/versions/3.0-10/topics/Predict
https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/predict

type="response"속성: 예측 결과를 0-1사이의 확률값으로 예측치를 얻기 위해서 지정

모델 평가를 위해서 예측치가 확률값으로 제공되기 때문에 이를 이항형으로 변환하는 과정이 필요 → Ifelse()함수를 이용하여 예측치의 벡터변수(pred)를 입력으로 이항형의 벡터변수(result_pred)를 생성

6단계: 모델평가 – 분류정확도 계산

table(result_pred, test$RainTomorrow)

* 분류정확도는 데이터에 따라 상이

7단계: ROC(Receiver Operating Characteristic) Curve를 이용한 모델 평가
ROCR 패키지 설치

```
pr <- prediction(pred, test$RainTomorrow)
prf <- performance(pr, measure = "tpr", x.maeasure = "fpr")
plot(prf)
```

prediction()함수
https://www.rdocumentation.org/packages/ROCR/versions/1.0-11/topics/prediction

performance()함수
https://www.rdocumentation.org/packages/ROCR/versions/1.0-6/topics/performance

ROC curve에서 왼쪽 상단의 계단 모양의 빈 공백만큼이 분류정확도에서 오분류(missing)를 나타낸다.

통계기반 데이터분석 교과교재 PP42-45 실습