Logistic Regression with Rattle

한신대학교 응용통계학과 4학년 201452024 박상희

목차

1. 데이터 소개

Mroz 데이터는 결혼한 백인 여성이 직업을 갖는데 어떤 요인들이 영향을 미치는지에 대해 조사한 데이터이다. 총 8개의 변수가 있으며, 총 783개의 관찰값을 가지고 있다.

변수 설명

변수명	설명	특성	예시
lfp	현재 직업 유무(Target)	Factor w/ 2 levels "no","yes"	2 2 2 2 2 2 2 2 2 2 ...
k5	5세 이하 자녀의 수	int	1 0 1 0 1 0 0 0 0 0 ...
k618	6 ~ 18세의 자녀의 수	int	0 2 3 3 2 0 2 0 2 2 ...
age	백인 여성의 나이	int	32 30 35 34 31 54 37 54 48 39 ...
wc	여성의 대학 졸업 여부	Factor w/ 2 levels "no","yes"	1 1 1 1 2 1 2 1 1 1 ...
hc	남편의 대학 졸업 여부	Factor w/ 2 levels "no","yes"	1 1 1 1 1 1 1 1 1 1 ...
lwg	여성의 예상 임금률	num	1.2102 0.3285 1.5141 0.0921 1.5243 ...
inc	여성의 소득을 제외한 가족의 전체 소득	num	10.9 19.5 12 6.8 20.1 ...

데이터 예시

	lfp	k5	k618	age	wc	hc	lwg	inc
1	yes	1	0	32	no	no	1.2101647	10.910001
2	yes	0	2	30	no	no	0.3285041	19.500000
3	yes	1	3	35	no	no	1.5141279	12.039999
4	yes	0	3	34	no	no	0.0921151	6.800000
5	yes	1	2	31	yes	no	1.5242802	20.100000
6	yes	0	0	54	no	no	1.5564855	9.859000
7	yes	0	2	37	yes	no	2.1202636	9.152000
8	yes	0	0	54	no	no	2.0596387	10.900001
9	yes	0	2	48	no	no	0.7543364	17.305000
10	yes	0	2	39	no	no	1.5448993	12.925000
11	yes	0	1	33	no	no	1.4019157	24.300001
12	yes	0	1	42	no	yes	1.5242802	19.699999
13	yes	1	2	30	no	yes	0.7339690	15.000000
14	yes	0	2	43	no	no	0.8183691	14.600000
15	yes	0	1	43	no	yes	1.3028187	24.631000
16	yes	0	3	35	no	no	0.2980447	17.531000

2. 전체 데이터에 대한 모형 적합 및 오분류표

 # 제시된 회귀 모형

$$X'\beta = \beta_0 + \beta_1 inc + \beta_2 wc + \beta_3 lwg + \beta_4 lwg^2 + \beta_5 age + \beta_6 k5 + \beta_7 k618$$

 # 적합된 회귀 모형

$$X'\beta = 7.45568 - 0.02944 inc + 0.22377 wc[yes] - 8.54061 lwg + 4.85566 lwg^2 - 0.07922 age - 1.84950 k5 - 0.08429 k618$$

 # 오분류표

Full data (obs = 753)		Predicted	
		No	Yes
Actual	No	243	82
	Yes	111	317

3. Stepwise에 의한 모형 선택 및 선택된 모형의 오분류표

 # 단계적 변수 선택법에 의한 회귀 모형

$$X'\beta = \beta_0 + \beta_1 k5 + \beta_2 age + \beta_3 lwg + \beta_4 inc + \beta_5 wc[yes]$$

 # 적합된 회귀 모형

$$X'\beta = 2.841505 - 1.536700 k5 - 0.057930 age + 0.543253 lwg - 0.030853 inc + 0.992665 wc[yes]$$

 # 오분류표

Full data (obs = 753)		Predicted	
		No	Yes
Actual	No	190	135
	Yes	98	330

4. 데이터 분할과 Test 데이터에 대한 오분류표

위의 2개의 모델에 대하여 7:3의 비율로 Train 데이터와 Test 데이터를 분리한 다음, Train 데이터로만 모형을 적합시킨 후, Tarin 데이터와 Test 데이터에 대한 오분류표를 계산

 # 오분류표

Train (obs = 527)		Predicted	
		No	Yes
Actual	No	179	55
	Yes	72	221

[Model_1]

Train (obs = 527)		Predicted	
		No	Yes
Actual	No	139	95
	Yes	69	224

[Model_2]

Test (obs = 226)		Predicted	
		No	Yes
Actual	No	64	27
	Yes	39	96

[Model_1]

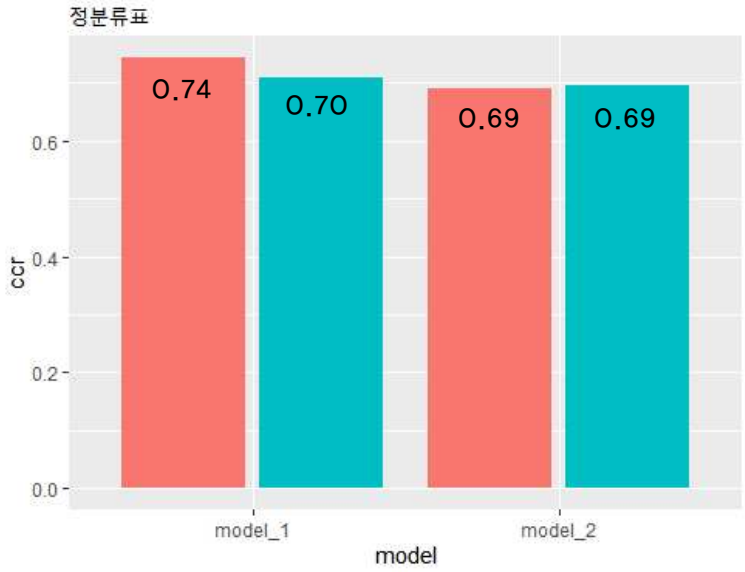
Test (obs = 226)		Predicted	
		No	Yes
Actual	No	51	40
	Yes	29	106

[Model_2]

5. 각각의 모형에 따른 오분류표 비교

4개의 모형에 대한 정분류율

	Full + Model1	Full + Model 2	Test + Model 1	Test + Model 2
정분류율	0.7436919	0.690571	0.7079646	0.6946903



그래프를 보면 전체 데이터를 사용했을 때보다 데이터를 분할해서 사용했을 때 정확도가 더 낮아지거나 비슷했다. 이러한 현상은 오버피팅이다. 오버피팅(Overfitting)이란 모델이 학습한 데이터에 너무 복잡하게 생성되어, 새로운 데이터가 들어왔을 때 잘 예측을 하지 못하는 현상이다.

2번 모델의 경우 오히려 예측 정확도는 조금 올라간 것을 알 수 있는데, 이는 2번 모델을 만들 때 정확도를 기준으로 생성한 것이 아닌 통계적 모형에 의한 단계별 선택법에 의해 변수 선택을 한 모형이기 때문이라고 판단되어 진다.

6. R 프로그래밍 코드

```
# 데이터 확인
library(car)
mroz <- Mroz
str(mroz)
summary(mroz)
head(mroz)

# 첫번째 모형
library(tidyverse)
model_1 <- mroz %>% dplyr::mutate(lwg_2 = I(lwg^2)) %>% dplyr::select(lfp, inc, wc, lwg, lwg_2, age, k5, k618)

# 단계별 선택법
library(MASS)
fit_full <- glm(lfp ~ ., family="binomial", data=Mroz)     # FULL Model
fit <- glm(lfp ~ 1, family="binomial", data=Mroz)          # MAIN Model
addterm(fit, fit_full, test="Chisq")                       # LRT 통계량이 가장 큰 값 K5
fit <- update(fit, . ~ . + k5)                             # MAIN Model 에 K5 변수 추가
dropterm(fit, test="Chisq")                                # MAIN Model 에서 K5에 대한 유의성 검정 실시
addterm(fit, fit_full, test="Chisq")                       # LRT 통계량이 가장 큰 값  age
fit <- update(fit, . ~ . + age)                            # MAIN Model 에 age 변수 추가
dropterm(fit, test="Chisq")                                # MAIN Model 에서 age에 대한 유의성 검정 실시
addterm(fit, fit_full, test="Chisq")                       # LRT 통계량이 가장 큰 값 lwg
fit <- update(fit, . ~ . + lwg)                            # MAIN Model 에 lwg 변수 추가
dropterm(fit, test="Chisq")                                # MAIN Model 에서 lwg에 대한 유의성 검정 실시
addterm(fit, fit_full, test="Chisq")                       # LRT 통계량이 가장 큰 값 inc
fit <- update(fit, . ~ . + inc)                            # MAIN Model 에 inc 변수 추가
dropterm(fit, test="Chisq")                                # MAIN Model 에서 inc에 대한 유의성 검정 실시
addterm(fit, fit_full, test="Chisq")                       # LRT 통계량이 가장 큰 값 wc
fit <- update(fit, . ~ . + wc)                             # MAIN Model 에 wc 변수 추가
dropterm(fit, test="Chisq")                                # MAIN Model 에서 wc에 대한 유의성 검정 실시
addterm(fit, fit_full, test="Chisq")                       # 단계별 선택법 종료

# 두번째 모형
model_2 <- mroz %>% dplyr::select(lfp, k5, age, lwg, inc, wc)

# rattle 실행
library(rattle)
rattle()

# 4개의 모델에 대한 정분류율
data.frame(model = c("model_1", "model_2", "model_1", "model_2"),
           data = c("Full","Full","Test","Test"),
           ccr = c(0.7436919, 0.690571, 0.7079646, 0.6946903)) %>%
  ggplot(aes(x=model, y=ccr, fill=data)) +
  geom_bar(stat="identity", position="dodge2") +
  labs(title="정분류표") +
  theme(legend.position = "none")
```