REPORT

통계기반 데이터 분석

로지스틱 회귀분석을 활용한
엔진, 연비, 변속기의 상관관계에 대하여
- mtcars 데이터셋을 바탕으로

2023.03.20

B1 팀

김예지, 서영석, 이현빈, 전국림

<목차>

1. 서론

1) 데이터 분석 배경

산업혁명부터 이어져 온 인류의 기술 발전은 운송 수단의 발전을 대표적으로 포함한다. 그 중 가장 친숙하게 접할 수 있는 운송수단이 자동차이다. 이를 증명하듯, 자동차는 운송수단 중 가장 높은 사용량과 다양한 기종을 지니고 있다. 또한 자동차의 구성품에 따른 성능을 확인하고자 하는 움직임이 자연스럽게 나타난다. 이 보고서에서는 자동차의 요소 중 핵심이라고 할 수 있는 엔진, 연비, 변속기 종류의 상관관계에 대해 파악하고자 한다. 이를 통해 차종 선택 및 신차 개발 시 유의 사항 마련을 위한 데이터적 근거 기반을 형성하고자 한다.

2) 데이터 분석 설명

이 보고서에서는 R을 활용하여 자동차의 엔진, 연비, 변속기의 상관관계에 대한 데이터를 분석할 예정이다. 데이터셋은 일관성을 위해 R의 내장 데이터셋인 mtcars 데이터셋으로 제한한다.

해당 데이터는 32개의 차종을 표본으로 한다. 또한 데이터셋 상의 변수명의 의미를 서론에서 먼저 밝혀, 보고서 내용 및 결과에 대한 이해를 돕고자 한다. 'am'은 변속기 종류를 의미한다. 입력 데이터 값은 0과 1이 있으며, 이는 각각 automatic(자동), manual(수동)을 의미한다. 'vs'는 엔진 종류에 대한 변수이다. 입력 데이터 값은 0과 1로 동일하게 들어가 있다. 이는 각각 V-shaped, straight를 의미한다.

3) 데이터 정제

본론에 들어가기에 앞서, 데이터셋을 가져오는 과정에 대해 먼저 밝히고자 한다. 먼저 "data(mtcars)"코드를 통해 mtcars 데이터셋을 로드(load)한다. 데이터셋에 대한 내용은 "?mtcars"를 통해 확인할 수 있다. 이후 중심적으로 활용할 데이터인, mtcars 데이터를 "mtcars_df"변수로 지정한다. 해당 부분에 대한 코드는 아래와 동일하다. 또한, 데이터에 대한 기본적인 이해를 돕기 위해 기술 통계량 및 데이터 구성을 summary 함수와 dim 함수를 통해 확인하였다.

```
# 데이터 가져오기
data(mtcars)
?mtcars
#mtcars_df 담아서 데이터 확인하기
mtcars_df <- mtcars
head(mtcars_df)
```

```
summary(mtcars_df)
nrow(mtcars_df)
ncol(mtcars_df)
```

mtcars_df 에 담긴 데이터에는 총 32 개 자동차의 정보가 담겨져 있으며, gear, mpg 등 11 개의 특성을 칼럼으로 담고 있다. 해당 코드의 결과는 아래와 동일하다.

```
#mtcars 데이터 불러오기
> mtcars_df <- mtcars
> nrow(mtcars_df)
[1] 32
> ncol(mtcars_df)
[1] 11

> head(raw_data)
```

	mpg	cyl	disp	hp	drat
Mazda RX4	21.0	6	160	110	3.90
Mazda RX4 Wag	21.0	6	160	110	3.90
Datsun 710	22.8	4	108	93	3.85
Hornet 4 Drive	21.4	6	258	110	3.08
Hornet Sportabout	18.7	8	360	175.3	15
Valiant	18.1	6	225	105	2.76

	wt	qsec	vs	am	gear	crab
Mazda RX4	2.620	16.46	0	1	4	4
Mazda RX4 Wag	2.875	17.02	0	1	4	4
Datsun 710	2.320	18.61	1	1	4	1
Hornet 4 Drive	3.215	19.44	1	0	3	1
Hornet Sportabout	3.440	17.02	0	0	3	2
Valiant	3.460	20.22	1	0	3	1

또한 보다 명확한 이해를 위해 mtcars 데이터셋에 있는 변수들에 대한 정의를 확립하고자 한다. 정의는 아래 표와 같다.

mtcars 변수		
mpg	갤런당 마일(연비)	수치형
cyl	실린더 갯수	수치형
disp	배기량	수치형
hp	마력	수치형
drat	리얼 액슬 비율	수치형
wt	무게	수치형
qsec	1/4 마일에 도달하는 시간	수치형
vs	엔진 종류 (0 = V-shaped, 1 = straight)	범주형
am	변속기 종류 (0=automatic, 1=manual)	범주형
gear	전진 기어 갯수	수치형
carb	기화기 갯수	수치형

2. 본론

1) 엔진(vs)변수 중심의 데이터 분석

이 보고서에서는 자동차 구성 요소의 상관 관계에 대하여 확인하고자 한다. 그 중에서도 엔진, 연비, 변속기 종류의 상관관계를 파악할 예정이다. 따라서 엔진(vs) 변수를 종속변수로 설정하고 나머지 변수를 독립변수로 설정했다. 또한, 로지스틱 회귀분석을 통해 결과값을 예측하고자 한다.

3 개의 요인의 상관 관계를 확인하기 위해서는 모든 요인을 포함한 분석값이 존재해야 한다. 따라서 엔진(vs)변수를 중심으로 두 개의 요인(칼럼)에 대한 로지스틱 회귀분석을, glm 함수를 활용해 실시했다. 이때, family=binomial 로 설정해주면 이항 로지스틱 분류가 가능하며, 이를 활용했다.

```
fx <- vs~mpg+am
mtcars_model <- glm(fx+am, data = mtcars_df,family = 'binomial', na.action=na.omit)
mtcars_model
summary(mtcars_model)
```

위의 코드에 대한 결과 값은 아래 사진과 동일하게 출력된다.

Call:
glm(formula = fx + am, family = binomial, data = mtcars)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.05888	-0.44544	-0.08765	0.33335	1.68405

Coefficients:

	Estimate	Std.Error	z value	Pr(>|z|)
(Intercept)	-12.7051	4.6252	-2.747	0.00602 **
mpg	0.6809	0.2524	2.698	0.00697 **
am	-3.0073	1.5995	-1.880	0.06009 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 43.860 on 31 degrees of freedom
Residual deviance: 20.646 on 29 degrees of freedom
AIC: 26.646

해당 결과를 통해 연비와 변속기에 따른 엔진 성능에 대한 예측이 가능하다. mpg(연비)의 회귀계수가 0.6809 이기 때문에 mpg 가 한 단위 증가하면 vs=1 일 오즈(odds)가 exp(0.6809) ≒1.98(98%) 증가한다. 또한 am(변속기)의 회귀계수는 -3.0073 이다. 따라서 am 이 한 단위 증가하면 (여기서는 0 에서 1 로 증가하는 것이므로, 자동에서 수동으로 변환되는 것을 의미한다.), vs=1 일 오즈가 exp(-3.0073) ≒ 0.05 가 된다. 이는 변속기가 수동인 경우, 자동에 비해 vs=1 인 오즈가 95% 감소함을 의미한다.

여기서 오즈(odds)는 승산을 의미한다. 승산은 기준이 되는 변수가 한 단위 증가할 경우, 확률이 변화하는 비율을 의미한다. 즉, 집단 1 에 속하는 확률에서 집단 0 에 속하는 확률을 나눈 비율이다.

2) 로지스틱 회귀 분석 시행

2-1) 변수 선택법 적용

앞서 분석한 결과 값과 관련한 로지스틱 회귀 모델을 정리하고자 한다. 이에 따라 유의한 변수들만 선택하기 위해 후진제거법 및 이탈도를 확인했다. 해당 코드는 아래와 같다.

```
# 1) 회귀식 생성하기
func <- glm(fx, data = mtcars_df,family = 'binomial', na.action=na.omit)
# 2) 후진제거법을 활용한 변수 선택
selectCar <- step(func, direction = "backward")
# 3) 이탈도를 확인
anova(func, test="Chisq")
```

위의 코드에 대한 결과 값은 아래 사진과 동일하게 출력된다.

Start: AIC=26.65

vs ~ mpg + am

	Df	Deviance	AIC
<none>		20.646	26.646
-am	1	25.533	29.533
-mpg	1	42.953	46.953

후진제거법 사용

Analysis of Deviance Table

Model: binomial, link: logit

Response: vs

Terms added sequentially (first to last)

	Df	Deviance	Resid.df	resid.Dev	Pr(>Chi)
<none>			31	43.860	
-am	1	18.327	30	25.533	1.861e-05 ***
-mpg	1	4.887	29	20.646	0.02706 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

이탈도 확인

후진 제거법을 진행했을 때, am 과 mpg 의 AIC 가 모두 기준 AIC 보다 적으므로 두 변수 모두 최적의 예측변수이다. 더 나아가 모든 변수가 유의한지 확인하기 위해 이탈도를 계산한 결과는, mpg 와 am 모두 Pr(>Chi)가 0.05 이하이다. 따라서 통계적으로 유의하다고 할 수 있다.

2-2) 예측치 설정

회귀분석을 한 데이터인 mtcars_model 변수와 새로운 데이터인 mtcars_df(초기 데이터)를 input 시켜 새로운 데이터에 대한 예측된 값을 구하고자 한다. 이를 위해 predict()함수를 활용했으며, type="response" 파라미터를 추가해 '확률'의 꼴로 반환했다. 이를 의미하는 코드와 결과는 아래와 동일하다.

```
pred <- predict(mtcars_model, newdata = mtcars_df, type = "response")
pred
```

Mazda RX4	Mazda RX4 Wag	Datsun 710
0.195752093	0.195752093	0.453286237
Hornet 4 Drive	**Hornet Sportabout**	**Valiant**
0.866062468	0.507024063	0.406017350
Duster 360	**Merc 240D**	**Merc 230**
0.048894865	0.980340613	0.943740285
Merc 280	**Merc 280C**	**Merc 450SE**
0.591110583	0.357844862	0.176823422
Merc 450SL	**Merc 450SLC**	**Cadillac Fleetwood**
0.283901447	0.086659370	0.003598826
Lincoln Continental	**Chrysler Imperial**	**Fiat 128**

0.003598826	0.063234437	0.998255316
Honda Civic	**Toyota Corolla**	**Toyota Corona**
0.993224169	0.999371042	0.873766034
Dodge Challenger	**AMC Javelin**	**Camaro Z28**
0.104252045	0.086659370	0.025360551
Pontiac Firebird	**Fiat X1-9**	**Porsche 914-2**
0.591110583	0.946684602	0.879906401
Lotus Europa	**Ford Pantera L**	**Ferrari Dino**
0.993224169	0.007006782	0.091267566
Maserati Bora	**Volvo 142E**	
0.004075891	0.242193639	

이후, 컷오프를 0.5 로 단순 가정하여 분류값을 생성한다. 0.5 이상일 경우 1, 0.5 이하일 경우 0 으로 반환한다. 이를 적용한 코드와 결과값은 아래와 동일하다.

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

Mazda RX4	Mazda RX4 Wag	Datsun 710
0	0	0
Hornet 4 Drive	**Hornet Sportabout**	**Valiant**
1	1	0
Duster 360	**Merc 240D**	**Merc 230**
0	1	1
Merc 280	**Merc 280C**	**Merc 450SE**
1	0	0
Merc 450SL	**Merc 450SLC**	**Cadillac Fleetwood**
0	0	0
Lincoln Continental	**Chrysler Imperial**	**Fiat 128**
0	0	1
Honda Civic	**Toyota Corolla**	**Toyota Corona**
1	1	1
Dodge Challenger	**AMC Javelin**	**Camaro Z28**
0	0	0
Pontiac Firebird	**Fiat X1-9**	**Porsche 914-2**
1	1	1

Lotus Europa	Ford Pantera L	Ferrari Dino
1	0	0
Maserati Bora	**Volvo 142E**	
0	0	

result_pred	0	1
	19	13

3) 로지스틱 회귀모델 요약 및 평가

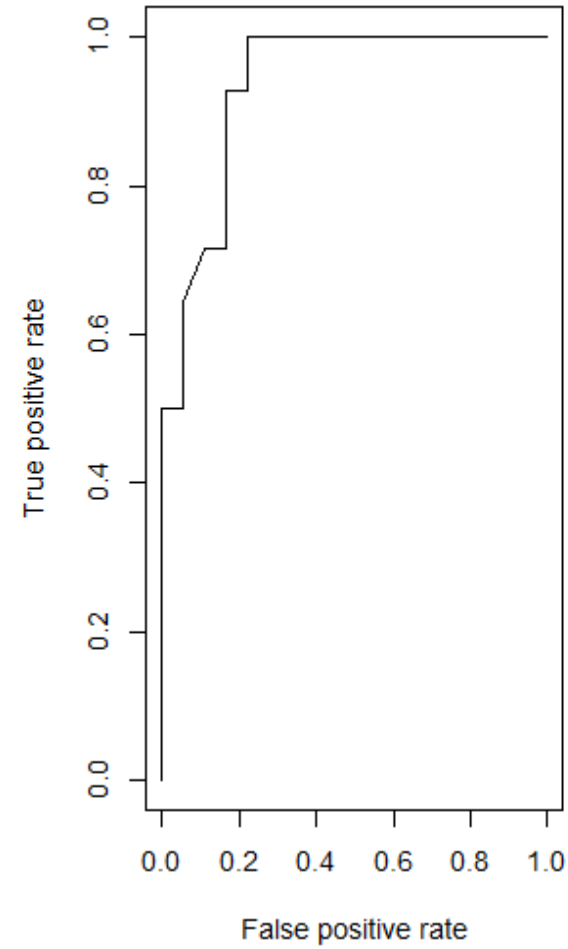
3-1) 회귀 모델 요약

table()함수로 예측 성능이 얼마나 좋은지 수치로 표현했으며 Performance()를 통해 ROC curve 로 나타냈고, 이를 바탕으로 분류 정확도를 도출했다.

```
table(result_pred, mtcars_df$vs)
```

result_pred	0	1
0	15	4
1	3	10

테이블에 따르면, vs 가 올바로 예측된 경우는 25 개(10+15), 잘못 예측된 경우는 7(3+4)개이다.




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

또한, ROC curve 에서 왼쪽 상단의 계단 모양의 빈 공백만큼이 분류정확도에서 오분류(missing)를 나타낸다. 또한 분류 정확도는 테이블 함수의 결과 값을 활용하여, 테이블의 "대각석의 합/전체 합"으로 도출할 수 있다. 여기서는 "(15+10) / nrow(mtcars_df)"로 계산되며, 0.78125 가 나왔다. 즉 이 모델은 약 78.125 퍼센트의 신뢰 모델을 가지고 있다고 평가할 수 있다.

3-2) 회귀 분석 식

x1 은 mpg 값에 해당하는 변수이고, x2 는 am 에 해당하는 변수, p 는 vs 가 1 일 확률이라 할 때 vs 를 종속변수로 x1 과 x2 에 관한 로지스틱 회귀방정식은 다음과 같다.

ln(p/1-p) = -12.7051 + (0.6809*x1) + (-3.0073*x2)

상관관계를 중심으로 회귀 방정식을 생성하면, 위와 동일한 방정식이 도출된다. 이를 토대로 가설에 대한 결과값을 확인할 수 있다.

3. 분석 결과 및 결론

1) 데이터 분석 결과 및 결론 도출

엔진, 연비, 변속기 종류의 상관관계에 대해 확인했다. 정리를 하면, mpg 와 vs 는 양의 관계를, am 과 vs 는 음의 관계를 가진다. mpg(연비)가 한 단위 증가하면 'vs(엔진)가 한단위 증가할 가능성'이 98% 증가한다. 또한 am(변속기)이 한 단위 증가하면, 즉 자동에서 수동으로 변환되면, 'vs 가 한단위 증가할 가능성'이 0.05 가 된다. 이는 변속기가 수동인 경우, 자동에 비해 vs=1 인 오즈가 95% 감소함을 의미한다. 또한 이 모델은 약 78.125 퍼센트의 신뢰 모델을 가지고 있다.

참고 자료

1) 데이터셋: R 내 기본 장착 데이터셋, mtcars