채널 G
2016125077 최재혁
2016126049 박희재

데이터 사이언스

- 1주차 용돈 예측모델

과제

- 과제 설명

  현재 내 상황(조건)에서 엄마에게 용돈을 요구했을 시 어떤 결과를 얻어 낼지 판단하는 모델을 생성

- 과제 목표

  과거 데이터들의 패턴은 어떤 지, 데이터들 간의 연관성은 얼마나 있는지 파악을 하여 최대한 정확한 모델을 생성해 보겠습니다

- 사용 기법

  분석용 언어 : R

  분석 기법 : 다중선형회귀분석

회귀분석

- 정의

통계학에서 사용하는 자료 분석 방법 중 하나로, 간략히 표현해 여러 자료들 간의 관계성을 수학적으로 추정, 설명한다 (출처:namu.wiki)

하나의 종속(목표)변수

여러개의 독립변수

- 특징

1. 종속변수와 독립변수 간에 선형관계가 존재하는지 알 수 있다
2. 종속변수에 영향을 미치는 독립변수가 유의 한지와 영향력의 정도를 알 수 있다
3. 추정된 회귀모형을 통해 종속변수의 예측치를 알 수 있다

하나의 종속(목표)변수

하나의 독립변수

훈련 및 테스트 데이터

- 훈련데이터(30개)

엄마의기분 시험점수 기온 코로나발생여부 아빠의 보너스 결과

	A	B	C	D	E	F
1	-1	80	23.5	0	200	0
2	1	85	27.5	1	100	1
3	0	90	29.4	0	150	1
4	0	75	30	0	100	0
5	1	95	23	0	200	1
6	-1	65	27.5	1	150	0
7	0	100	21	1	200	1
8	0	85	27	1	100	1
9	1	85	27	0	150	1
10	-1	50	21	0	200	0
11	-1	90	26.1	1	150	0
12	1	70	23	0	200	1
13	-1	30	28	0	200	0
14	1	60	24	0	150	1
15	1	30	27.6	1	100	0
16	1	70	24	0	150	1
17	1	88	23.5	1	100	1
18	1	85	21.3	0	150	1
19	1	90	25	1	200	1

training1

- 테스트데이터(30개)

엄마의기분 시험점수 아빠의 보너스 결과

	A	B	C	D
1	-1	75	100	0
2	1	85	150	1
3	0	90	120	1
4	0	60	100	0
5	1	95	200	1
6	-1	70	150	0
7	0	90	200	1
8	0	95	100	1
9	1	75	150	1
10	-1	55	200	0
11	-1	90	150	1
12	1	70	100	1
13	-1	20	150	0
14	1	55	150	1
15	1	28	100	0
16	1	71	150	1
17	1	85	120	1
18	1	87	130	1
19	1	90	200	1

testing1

데이터 가져오기

입력

```
pre_dat × | homework1.R* ×
Source on Save                                   Run | Source
1  df <- read.csv(file = "C:/Users/user/Desktop/3학년2학기/데이터사이언스/과제/training1.csv",header=TRUE, fileEncoding = "UCS-2LE")
2  colnames(df) <- c("mode","score","tem","cov","bon","ans")
3  df
4
```

설명

R언어의 read.csv함수를 이용하여 file에 저장된 훈련데이터를 가져와서 속성들의 이름을 바꾸어 주었습니다

출력

```
Console | Terminal × | Jobs ×
C:/Users/user/Desktop/3학년2학기/데이터사이언스/과제/

[Workspace loaded from C:/Users/user/Desktop/3?숇퀎2?숇린/?겟씠?겟궗?댁뼵??怨쇱젣/.RData]

> df <- read.csv(file = "C:/Users/user/Desktop/3학년2학기/데이터사이언스/과제/training1.csv",header=TRUE, fileEncoding = "UCS-2LE")
> colnames(df) <- c("mode","score","tem","cov","bon","ans")
> df
   mode score  tem cov bon ans
1     1    85 27.5   1 100   1
2     0    90 29.4   0 150   1
3     0    75 30.0   0 100   0
4     1    95 23.0   0 200   1
5    -1    65 27.5   1 150   0
6     0   100 21.0   1 200   1
7     0    85 27.0   1 100   1
8     1    85 27.0   0 150   1
9    -1    50 21.0   0 200   0
10   -1    90 26.1   1 150   0
11    1    70 23.0   0 200   1
12   -1    30 28.0   0 200   0
13    1    60 24.0   0 150   1
14    1    30 27.6   1 100   0
15    1    70 24.0   0 150   1
16    1    88 23.5   1 100   1
17    1    85 21.3   0 150   1
18    1    90 25.0   1 200   1
19    0    70 28.2   0 150   0
20    0    75 24.3   1 100   0
21    0    80 28.0   1 200   0
22    0    90 26.0   1 150   1
```

데이터 요약 정리

입력

```
summary(df)
str(df)
```

설명

summary()함수를 이용하여 해당 데이터프레임에 대한 요약된 정보를 추출해 보았습니다
(Min:최소값, 1st Qu: 1분위수(25%), Median: 중간값, Mean:평균값, 3rd Qu:3분위수(75%), Max:최대값)

str()함수를 이용하여 데이터 구조, 변수 개수, 변수 명, 관찰치 개수, 관찰치를 확인하였습니다

출력

```
> summary(df)
      mode        score            tem             cov              bon             ans
 Min.   :-1   Min.   : 30.00   Min.   :20.50   Min.   :0.0000   Min.   :100.0   Min.   :0.0000
 1st Qu.:-1   1st Qu.: 70.00   1st Qu.:23.50   1st Qu.:0.0000   1st Qu.:150.0   1st Qu.:0.0000
 Median : 0   Median : 85.00   Median :26.00   Median :0.0000   Median :150.0   Median :1.0000
 Mean   : 0   Mean   : 77.17   Mean   :25.34   Mean   :0.4483   Mean   :155.2   Mean   :0.5172
 3rd Qu.: 1   3rd Qu.: 90.00   3rd Qu.:27.50   3rd Qu.:1.0000   3rd Qu.:200.0   3rd Qu.:1.0000
 Max.   : 1   Max.   :100.00   Max.   :30.00   Max.   :1.0000   Max.   :200.0   Max.   :1.0000
> str(df)
'data.frame':   29 obs. of  6 variables:
 $ mode : int  1 0 0 1 -1 0 0 1 -1 -1 ...
 $ score: int  85 90 75 95 65 100 85 85 50 90 ...
 $ tem  : num  27.5 29.4 30 23 27.5 21 27 27 21 26.1 ...
 $ cov  : int  1 0 0 0 1 1 1 0 0 1 ...
 $ bon  : int  100 150 100 200 150 200 100 150 200 150 ...
 $ ans  : int  1 1 0 1 0 1 1 1 0 0 ...
>
```

모델 생성

입력

```
6  model <- lm(ans~mode+score+tem+cov+bon,data = df)
7  summary(model)
8
```

8:1 (Top Level) R Script

설명 lm ()함수를 이용하여 회귀분석을 실시하였습니다. 이때 data= 는 위에서 가져온 train data이며 Ans가 종속 변수가 되고 입력변수들은 ~ 오른쪽에 순서대로 써줍니다.

출력

```
> model <- lm(ans~mode+score+tem+cov+bon,data = df)
> summary(model)

Call:
lm(formula = ans ~ mode + score + tem + cov + bon, data = df)
```

중요한 점

***찍혀 있는 변수들은 회귀식에 유의한 즉, 영향을 끼치는 변수로 알 수 있습니다. 또한 나머지 변수들의 p-value값이 0.05, 즉 유의수준 5%보다 작아 유의하지 않은 것을 확인할 수 있습니다.
또한 모델의 설명력은 수정된 결정계수인 (Adjusted R-squared: 0.6373)로 알 수 있습니다. 63.7% 정도로 이 회귀식이 전체 데이터를 설명한다고 볼 수 있습니다.
F통계량의 결과 유의수준 0.05 보다 작은 1.898e-05로 모형은 통계적으로 유의함을 알 수 있습니다.

단계적 변수선택

입력

```
6  model <- lm(ans~mode+score+tem+cov+bon,data = df)
7  summary(model)
8  model1 <- step(model,direction = "backward")
9
```

설명 단계적 변수 선택, 즉 유의한 변수들만 남도록 후진제거법(backward)를 선택하였습니다

벌점화선택기준 - AIC

출력

```
> model1 <- step(model,direction = "backward")
Start:  AIC=-63.35
ans ~ mode + score + tem + cov + bon

        Df Sum of Sq    RSS     AIC
- tem    1   0.03189 2.1894 -64.927
- cov    1   0.07783 2.2353 -64.325
- bon    1   0.08991 2.2474 -64.168
<none>               2.1575 -63.352
- score  1   1.42170 3.5792 -50.673
- mode   1   2.83068 4.9882 -41.047

Step:  AIC=-64.93
ans ~ mode + score + cov + bon

        Df Sum of Sq    RSS     AIC
- cov    1   0.08193 2.2713 -65.861
- bon    1   0.13688 2.3262 -65.168
<none>               2.1894 -64.927
- score  1   1.42011 3.6095 -52.428
- mode   1   2.90944 5.0988 -42.410

Step:  AIC=-65.86
ans ~ mode + score + bon

        Df Sum of Sq    RSS     AIC
<none>               2.2713 -65.861
- bon    1    0.2968 2.5681 -64.300
- score  1    1.3382 3.6095 -54.428
- mode   1    3.0359 5.3072 -43.249
>
```

최종 모델생성

입력

```
9  model2 <- lm(ans~mode+score+bon,data = df)
10 summary(model2)
11
```

11:1 (Top Level) R Script

설명

유용한 변수를 가지고 새로운 변수인 model2를 생성하였습니다
설명력은 0.6487 즉, 64.8%로 기존보다 미세하게 상승하였으며 3변수 모두 유의한 것으로 판단할 수 있습니다
(bon은 0.08274로 완벽하게 유의한 변수는 아니지만 0.05와 비슷하다고 판단을 하여 넣어 주었습니다. 더 완벽한 모델을 만들기 위해서는 삭제하는 것이 맞는 방법입니다)

출력

Console Terminal Jobs

C:/Users/user/Desktop/3학년2학기/데이터사이언스/과제/

```
> model2 <- lm(ans~mode+score+bon,data = df)
> summary(model2)

Call:
lm(formula = ans ~ mode + score + bon, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.67361 -0.20191  0.04293  0.14276  0.53678

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.848952   0.327472  -2.592  0.01569 *
mode         0.397110   0.068697   5.781 5.03e-06 ***
score        0.012242   0.003190   3.838  0.00075 ***
bon          0.002716   0.001503   1.807  0.08274 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3014 on 25 degrees of freedom
Multiple R-squared:  0.6863,    Adjusted R-squared:  0.6487
F-statistic: 18.24 on 3 and 25 DF,  p-value: 1.757e-06
```

회귀 계수 추출 과 회귀식

입력

```
model2 <- lm(ans~mode+score+bon,data = df)
summary(model2)
coef(model2)
```

설명 coef()함수를 이용하여 모델의 회귀계수를 추출하였습니다

출력

```
> coef(model2)
 (Intercept)         mode        score          bon
-0.848952397  0.397109889  0.012242062  0.002715974
```

회귀식

Y = 0.397109889*mode + 0.012242062*score + 0.002715974*bon -0.848952397

모델 예측 데이터 준비

입력

```
12 pre_dat <- read.csv(file = "C:/Users/user/Desktop/3학년2학기/데이터사이언스/과제/testing1.csv",header=TRUE, fileEncoding = "UCS-2LE")
13 pre_dat
14 colnames(pre_dat) = c("mode","score","bon","ans")
15 pre_dat
16
```

16:1 (Top Level) R Script

출력

```
> colnames(pre_dat) = c("mode","score","bon","ans")
> pre_dat
   mode score bon ans
1     1    85 150   1
2     0    90 120   1
3     0    60 100   0
4     1    95 200   1
5    -1    70 150   0
6     0    90 200   1
7     0    95 100   1
8     1    75 150   1
9    -1    55 200   0
10   -1    90 150   1
11    1    70 100   1
12   -1    20 150   0
13    1    55 150   1
14    1    28 100   0
15    1    71 150   1
16    1    85 120   1
17    1    87 130   1
18    1    90 200   1
19    0    69 150   0
20    0    75 100   0
21    0    80 200   1
22    0    90 150   1
23    0    95 100   1
24   -1    58 140   0
25   -1    70 200   0
26   -1    80 200   0
27   -1    85 150   0
28   -1    90 150   0
29   -1   100 150   1
>
```

모델 예측(적용)

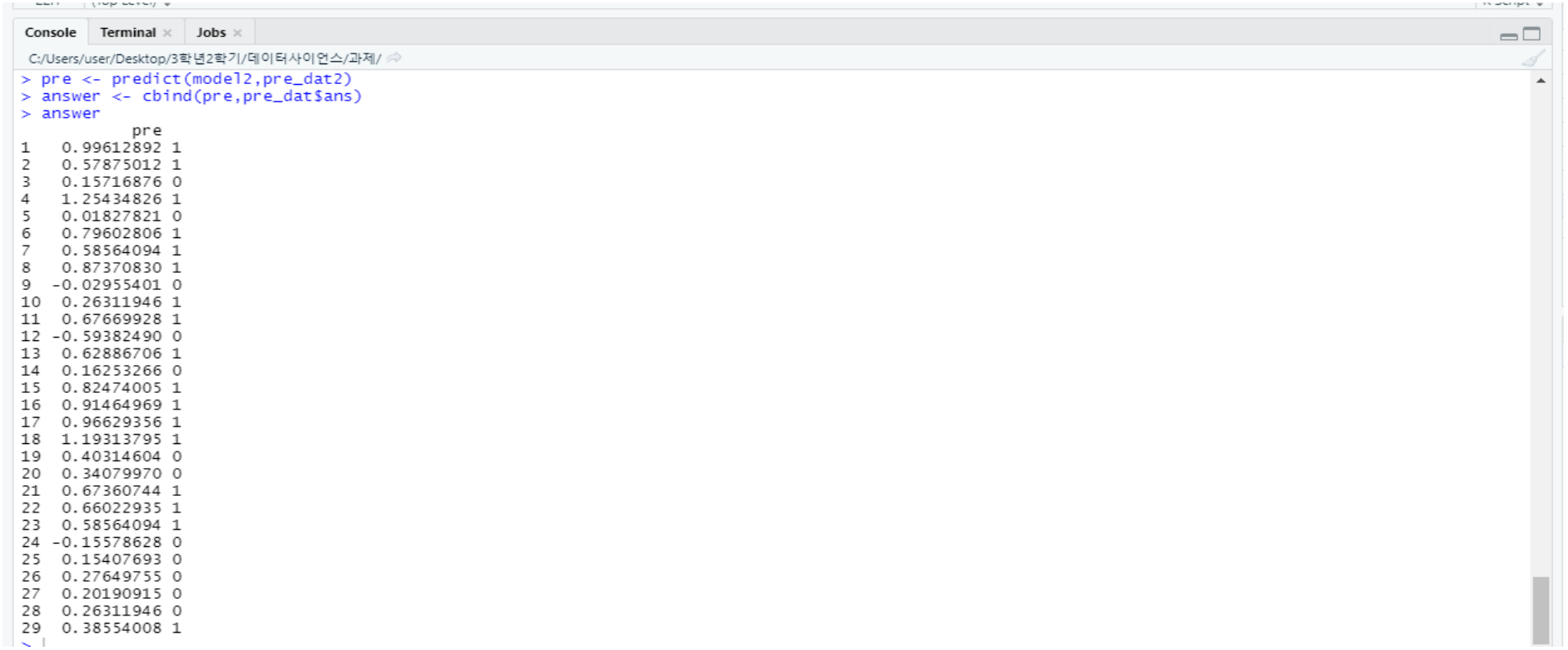
입력

```
16  library(sqldf)
17  pre_dat2 <- sqldf("select mode,score,bon from pre_dat")
18  pre_dat2
19  pre <- predict(model2,pre_dat2)
20  answer <- cbind(pre,pre_dat$ans)
21  answer
22
22:1  (Top Level)
```

설명

시험용 데이터 셋을 이전에 만든 model2에 집어 넣어서 예측을 시도하였습니다
왼쪽에 있는 값은 예측 값으로 점 추정한 결과입니다
오른쪽은 실제 저희가 정해준 결과 입니다

출력



임의의 값 넣어보기

설명 3가지의 임의 데이터를 넣어 보았습니다

입력

```
22 predict(model2,newdata = data.frame(mode=1,score=87,bon=130))
23 predict(model2,newdata = data.frame(mode=-1,score=100,bon=300))
24 predict(model2,newdata = data.frame(mode=0,score=60,bon=150))
25
```

17:56 (Top Level)

Console Terminal Jobs

C:/Users/user/Desktop/3학년2학기/데이터사이언스/과제/

출력

```
25  0.15407693 0
26  0.27649755 0
27  0.20190915 0
28  0.26311946 0
29  0.38554008 1
> predict(model2,newdata = data.frame(mode=1,score=87,bon=130))
        1
0.9662936
> predict(model2,newdata = data.frame(mode=-1,score=100,bon=300))
        1
0.7929362
> predict(model2,newdata = data.frame(mode=0,score=60,bon=150))
        1
0.2929675
>
```

최종 코드

pre_dat | homework1.R*

Source on Save | Run | Source

```r
df <- read.csv(file = "C:/Users/user/Desktop/3학년2학기/데이터사이언스/과제/training1.csv",header=TRUE, fileEncoding = "UCS-2LE")
colnames(df) <- c("mode","score","tem","cov","bon","ans")
df
summary(df)
str(df)
model <- lm(ans~mode+score+tem+cov+bon,data = df)
summary(model)
model1 <- step(model,direction = "backward")
model2 <- lm(ans~mode+score+bon,data = df)
summary(model2)
coef(model2)
pre_dat <- read.csv(file = "C:/Users/user/Desktop/3학년2학기/데이터사이언스/과제/testing1.csv",header=TRUE, fileEncoding = "UCS-2LE")
pre_dat
colnames(pre_dat) = c("mode","score","bon","ans")
pre_dat
library(sqldf)
pre_dat2 <- sqldf("select mode,score,bon from pre_dat")
pre_dat2
pre <- predict(model2,pre_dat2)
answer <- cbind(pre,pre_dat$ans)
answer
predict(model2,newdata = data.frame(mode=1,score=87,bon=130))
predict(model2,newdata = data.frame(mode=-1,score=100,bon=300))
predict(model2,newdata = data.frame(mode=0,score=60,bon=150))
```

17:56 (Top Level) R Script

결론 및 느낀점

각 변수가 유의해야 모델의 성능이 뛰어나다는 것을 알게 되면서 훈련 및 테스트 데이터의 중요성이 크다는 걸 알게 되었습니다

또한 회귀분석 모델을 실제로 만들어보고 예측 결과를 추출하는 과정에서 많은 것을 공부하게 되었습니다

R언어에 대한 구조와 다양한 함수를 찾아보니 분석에 많은 어려움을 덜어주고 자동화된 기능이 많다는 것을 알았습니다

그리고 하나의 예측에도 다양한 기법과 알고리즘을 사용하여 만들 수 있기 때문에 여러가지를 공부하고 적용하면서 폭넓은 분석을 해야 된다고 생각했습니다