Multiple Regression-test

Youngjin\_Lee

2022 2 22

Table of Contents

통계분석 관계검정 : 상관분석(Multiple regression) test 문제 K대학의 이교수는 온라인게임의 몰입(즐거움)에 영향을 주는 요인이 무엇인지를 연구하고자 한다. 문헌연구를 통해 도구(2문항), 보상(3문항), 정보(3문항), 디자인(4문항), 공동체(3문항), 몰입(4문항)을 요인으로 추출하고 설문지를 만들었다. 그렇다면 온라인게임 몰입에 영향을 주는 변수는 무엇이고, 어떤 변수가 온라인게임 몰입에 가장 큰 영향을 주는 변수인지 분석해 보자

# 1.기본 package 설정, library 로드

# 2.데이터 불러오기

mr\_tb <- read\_csv('data/MR.csv',   
 col\_names = TRUE,  
 locale=locale('ko', encoding='euc-kr'), # 한글  
 na=".")

## Rows: 175 Columns: 7  
## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## dbl (7): id, design, info, comm, op, fb, flow  
##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

str(mr\_tb)

## spec\_tbl\_df [175 x 7] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ id : num [1:175] 1 2 3 4 5 6 7 8 9 10 ...  
## $ design: num [1:175] 1.23 0.7 -0.19 1.2 0.14 -0.83 -0.45 -1.1 0.33 -1.67 ...  
## $ info : num [1:175] -1.46 -0.38 0.84 -1.21 -1.44 -1.45 0.29 -0.92 -1.47 -1.24 ...  
## $ comm : num [1:175] 1.67 -0.79 -1.12 -0.42 -1.48 -0.36 0.39 -3.02 -0.08 -0.47 ...  
## $ op : num [1:175] 0.76 -0.31 0.06 1.8 0.2 1.55 0.62 -0.94 0 -1.99 ...  
## $ fb : num [1:175] 1.22 1.29 -0.32 -0.27 0.68 -0.47 0.57 -0.21 0.08 1.46 ...  
## $ flow : num [1:175] 2.06 -0.34 0.75 -2.2 -0.18 -0.34 0.75 -3.29 -0.34 -1.11 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. id = col\_double(),  
## .. design = col\_double(),  
## .. info = col\_double(),  
## .. comm = col\_double(),  
## .. op = col\_double(),  
## .. fb = col\_double(),  
## .. flow = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

mr\_tb

## # A tibble: 175 x 7  
## id design info comm op fb flow  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 1.23 -1.46 1.67 0.76 1.22 2.06  
## 2 2 0.7 -0.38 -0.79 -0.31 1.29 -0.34  
## 3 3 -0.19 0.84 -1.12 0.06 -0.32 0.75  
## 4 4 1.2 -1.21 -0.42 1.8 -0.27 -2.2   
## 5 5 0.14 -1.44 -1.48 0.2 0.68 -0.18  
## 6 6 -0.83 -1.45 -0.36 1.55 -0.47 -0.34  
## 7 7 -0.45 0.29 0.39 0.62 0.57 0.75  
## 8 8 -1.1 -0.92 -3.02 -0.94 -0.21 -3.29  
## 9 9 0.33 -1.47 -0.08 0 0.08 -0.34  
## 10 10 -1.67 -1.24 -0.47 -1.99 1.46 -1.11  
## # ... with 165 more rows

# 3.기본통계치 확인

skim(mr\_tb)

Data summary

|  |  |
| --- | --- |
| Name | mr\_tb |
| Number of rows | 175 |
| Number of columns | 7 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| numeric | 7 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| id | 0 | 1 | 88 | 50.66 | 1.00 | 44.50 | 88.00 | 131.50 | 175.00 | ▇▇▇▇▇ |
| design | 0 | 1 | 0 | 1.00 | -3.01 | -0.58 | 0.03 | 0.66 | 2.56 | ▁▃▇▅▁ |
| info | 0 | 1 | 0 | 1.00 | -3.21 | -0.56 | 0.09 | 0.64 | 2.96 | ▁▃▇▅▁ |
| comm | 0 | 1 | 0 | 1.00 | -3.34 | -0.66 | -0.06 | 0.72 | 2.72 | ▁▂▇▆▁ |
| op | 0 | 1 | 0 | 1.00 | -2.48 | -0.72 | -0.06 | 0.70 | 3.23 | ▂▅▇▃▁ |
| fb | 0 | 1 | 0 | 1.00 | -2.29 | -0.68 | 0.02 | 0.69 | 3.28 | ▂▇▇▂▁ |
| flow | 0 | 1 | 0 | 1.00 | -3.29 | -0.53 | -0.18 | 0.75 | 2.61 | ▁▂▇▆▁ |

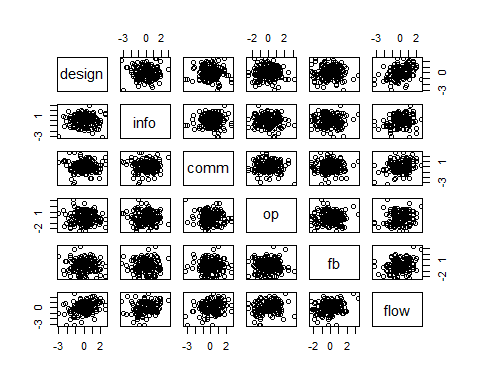
mr\_tb %>%  
 get\_summary\_stats()

## # A tibble: 7 x 13  
## variable n min max median q1 q3 iqr mad mean sd  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 comm 175 -3.34 2.72 -0.06 -0.665 0.72 1.38 1.02 0 1   
## 2 design 175 -3.01 2.56 0.03 -0.575 0.66 1.24 0.919 0 1   
## 3 fb 175 -2.29 3.28 0.02 -0.675 0.69 1.36 1.01 0 1   
## 4 flow 175 -3.29 2.61 -0.18 -0.535 0.75 1.28 1.05 -0.001 1   
## 5 id 175 1 175 88 44.5 132. 87 65.2 88 50.7  
## 6 info 175 -3.21 2.96 0.09 -0.56 0.64 1.2 0.875 0 1   
## 7 op 175 -2.48 3.23 -0.06 -0.725 0.695 1.42 1.07 0 1   
## # ... with 2 more variables: se <dbl>, ci <dbl>

mr\_tb <- mr\_tb %>%  
 select(-(id))

# 4.그래프 그리기(산점도)

pairs( ~ ., data=mr\_tb)



# 5.다중 회귀분석

전체변수 일괄입력 backward: 변수제거 forward: 변수추가 stepwise: backward와 forward 동시 AIC (Akaike information criterion), BIC (Bayesian …)

## 모델 생성

mr\_fit <- lm(flow ~ design+info+comm+op+fb,   
 data=mr\_tb) %>%  
 lm.beta() # 표준화 회귀계수

# ANOVA 분석

summary(mr\_fit)

##   
## Call:  
## lm(formula = flow ~ design + info + comm + op + fb, data = mr\_tb)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.51516 -0.40653 0.01118 0.51748 2.15898   
##   
## Coefficients:  
## Estimate Standardized Std. Error t value Pr(>|t|)   
## (Intercept) -0.0006925 0.0000000 0.0604381 -0.011 0.990871   
## design 0.3539020 0.3538603 0.0605969 5.840 2.61e-08 \*\*\*  
## info 0.2446373 0.2446131 0.0605957 4.037 8.19e-05 \*\*\*  
## comm 0.3312351 0.3311719 0.0606013 5.466 1.63e-07 \*\*\*  
## op 0.2085877 0.2085529 0.0605998 3.442 0.000728 \*\*\*  
## fb 0.2040364 0.2039402 0.0606183 3.366 0.000944 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7995 on 169 degrees of freedom  
## Multiple R-squared: 0.3796, Adjusted R-squared: 0.3612   
## F-statistic: 20.68 on 5 and 169 DF, p-value: 4.359e-16

tidy(mr\_fit)

## # A tibble: 6 x 6  
## term estimate std\_estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) -0.000693 0 0.0604 -0.0115 0.991   
## 2 design 0.354 0.354 0.0606 5.84 0.0000000261  
## 3 info 0.245 0.245 0.0606 4.04 0.0000819   
## 4 comm 0.331 0.331 0.0606 5.47 0.000000163   
## 5 op 0.209 0.209 0.0606 3.44 0.000728   
## 6 fb 0.204 0.204 0.0606 3.37 0.000944

# 회귀계수

tidy(mr\_fit, conf.int = TRUE) %>%  
 mutate\_if(is.numeric, round, 5)

## # A tibble: 6 x 8  
## term estimate std\_estimate std.error statistic p.value conf.low conf.high  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 (Interce~ -0.00069 0 0.0604 -0.0115 0.991 -0.119 0.119  
## 2 design 0.354 0.354 0.0606 5.84 0 0.234 0.473  
## 3 info 0.245 0.245 0.0606 4.04 0.00008 0.125 0.364  
## 4 comm 0.331 0.331 0.0606 5.47 0 0.212 0.451  
## 5 op 0.209 0.209 0.0606 3.44 0.00073 0.0889 0.328  
## 6 fb 0.204 0.204 0.0606 3.37 0.00094 0.0843 0.324

# 설명력R2

glance(mr\_fit)

## Warning: Tidiers for objects of class lm.beta are not maintained by the broom  
## team, and are only supported through the lm tidier method. Please be cautious in  
## interpreting and reporting broom output.

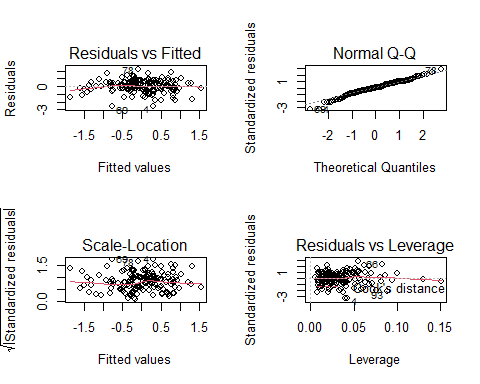
## # A tibble: 1 x 12  
## r.squared adj.r.squared sigma statistic p.value df logLik AIC BIC  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0.380 0.361 0.800 20.7 4.36e-16 5 -206. 426. 448.  
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>

# 6.회귀분석 가정 검정

등분산성: Scale-Location, ncvTest 정규성: Nomal Q-Q, shapiro.test 선형성: Residuals vs Fitted, 독립성: durbinWatsonTest 이상치검정 : Residuals vs Leverage(cook’s distance) 4/n-k-1

# 그림으로 가정 검정

opar <- par(no.readonly = TRUE)  
 par(mfrow=c(2,2))  
 plot(mr\_fit)



par(opar)

# 수치로 가정 검정

# 정규성 검정

shapiro\_test(mr\_fit$residuals)

## # A tibble: 1 x 3  
## variable statistic p.value  
## <chr> <dbl> <dbl>  
## 1 mr\_fit$residuals 0.986 0.0836

library(car)

## 필요한 패키지를 로딩중입니다: carData

##   
## 다음의 패키지를 부착합니다: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

## The following object is masked from 'package:purrr':  
##   
## some

# 등분산성 검정

car::ncvTest(mr\_fit)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 0.328302, Df = 1, p = 0.56666

# 독립성

car::durbinWatsonTest(mr\_fit)

## lag Autocorrelation D-W Statistic p-value  
## 1 -0.08283621 2.15399 0.28  
## Alternative hypothesis: rho != 0

#이상치 검정, sd, hat, d 통합검정

car::influencePlot(mr\_fit, id.method="identify")

## Warning in plot.window(...): "id.method"는 그래픽 매개변수가 아닙니다

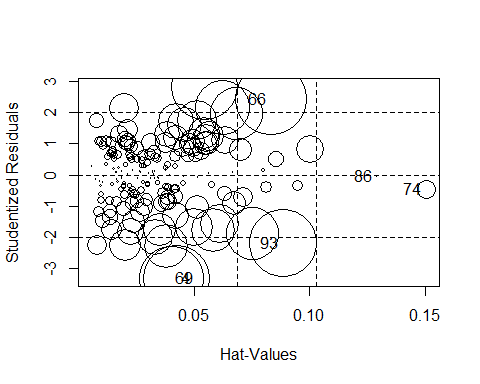
## Warning in plot.xy(xy, type, ...): "id.method"는 그래픽 매개변수가 아닙니다

## Warning in axis(side = side, at = at, labels = labels, ...): "id.method"는 그래  
## 픽 매개변수가 아닙니다  
  
## Warning in axis(side = side, at = at, labels = labels, ...): "id.method"는 그래  
## 픽 매개변수가 아닙니다

## Warning in box(...): "id.method"는 그래픽 매개변수가 아닙니다

## Warning in title(...): "id.method"는 그래픽 매개변수가 아닙니다

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "id.method"는 그래픽 매개  
## 변수가 아닙니다



## StudRes Hat CookD  
## 4 -3.29966948 0.04242307 7.594934e-02  
## 66 2.42797272 0.08320834 8.666266e-02  
## 69 -3.30344105 0.03995443 7.149907e-02  
## 74 -0.45755557 0.15038636 6.205260e-03  
## 86 -0.05172283 0.12938180 6.665446e-05  
## 93 -2.19253241 0.08868148 7.624799e-02

# 부록: 변수 제거 방법

backward: 변수제거 모든 변수 투입확인 제거된 변수있는지 확인 AIC값이 줄어들어야 의미가 있음

mr\_fit\_bk <- lm(flow ~ ., mr\_tb) %>%  
 lm.beta() # 표준화 회귀계수  
  
mr\_fit\_bk <- stats::step(mr\_fit\_bk, # stats:: 필수, 삭제시 에러발생  
 direction = "backward",   
 trace = T) # trace step별로 모두 보여주기

## Start: AIC=-72.42  
## flow ~ design + info + comm + op + fb  
##   
## Df Sum of Sq RSS AIC  
## <none> 108.03 -72.415  
## - fb 1 7.2422 115.27 -63.060  
## - op 1 7.5735 115.60 -62.558  
## - info 1 10.4189 118.45 -58.303  
## - comm 1 19.0972 127.13 -45.929  
## - design 1 21.8035 129.83 -42.243

# forward: 변수제거

mr\_fit\_fw <- lm(flow ~ 1, mr\_tb) %>%  
 lm.beta() # 표준화 회귀계수  
summary(mr\_fit\_fw)

##   
## Call:  
## lm(formula = flow ~ 1, data = mr\_tb)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.2893 -0.5343 -0.1793 0.7507 2.6107   
##   
## Coefficients:  
## Estimate Standardized Std. Error t value Pr(>|t|)  
## (Intercept) -0.0006857 0.0000000 0.0756201 -0.009 0.993  
##   
## Residual standard error: 1 on 174 degrees of freedom

mr\_fit\_fw <- stats::step(mr\_fit\_fw, # stats:: 필수, 삭제시 에러발생  
 direction = "forward",   
 scope =(flow ~ design+info+comm+op+fb), # "."대체시 에러  
 trace = T)

## Start: AIC=1.12  
## flow ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + design 1 21.7762 152.35 -20.2570  
## + comm 1 19.0888 155.04 -17.1969  
## + info 1 10.3936 163.73 -7.6474  
## + op 1 7.5677 166.56 -4.6528  
## + fb 1 7.2282 166.90 -4.2965  
## <none> 174.12 1.1231  
##   
## Step: AIC=-20.26  
## flow ~ design  
##   
## Df Sum of Sq RSS AIC  
## + comm 1 19.0950 133.25 -41.692  
## + info 1 10.4063 141.94 -30.638  
## + op 1 7.5707 144.78 -27.177  
## + fb 1 7.2335 145.12 -26.770  
## <none> 152.35 -20.257  
##   
## Step: AIC=-41.69  
## flow ~ design + comm  
##   
## Df Sum of Sq RSS AIC  
## + info 1 10.4046 122.85 -53.919  
## + op 1 7.5681 125.69 -49.925  
## + fb 1 7.2400 126.01 -49.469  
## <none> 133.25 -41.692  
##   
## Step: AIC=-53.92  
## flow ~ design + comm + info  
##   
## Df Sum of Sq RSS AIC  
## + op 1 7.5769 115.27 -63.060  
## + fb 1 7.2456 115.60 -62.558  
## <none> 122.85 -53.919  
##   
## Step: AIC=-63.06  
## flow ~ design + comm + info + op  
##   
## Df Sum of Sq RSS AIC  
## + fb 1 7.2422 108.03 -72.415  
## <none> 115.27 -63.060  
##   
## Step: AIC=-72.42  
## flow ~ design + comm + info + op + fb