Logistic Regression-test

Youngjin\_Lee

2022 2 25

Table of Contents

# 로지스틱 회귀분석(Logistic Regression)

통계분석 관계검정 : 로지스틱 회귀분석(Logistic Regression) t-test 문제 K기업의 인사담당인 이부장은 신체적 건강과 심리적건강, 조직몰입, 이직경험이 이직의도에 영향을 준다고 보고, 이들간의 인과관계를 연구하고자 한다

# 1.기본 package 설정, library 로드

# 2.데이터 불러오기

lr\_tb <- read\_csv('./data/LR.csv',   
 col\_names = TRUE,  
 locale=locale('ko', encoding='euc-kr'), # 한글  
 na=".") %>%  
 mutate\_if(is.character, as.factor) %>%  
 mutate(exp = factor(exp,  
 levels=c(0,1),  
 labels=c("No","Yes"))) %>%  
 mutate(chun = factor(chun,  
 levels=c(0:1),  
 labels=c("No","Yes")))

## Rows: 100 Columns: 5  
## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## dbl (5): phy, psy, cmmt, exp, chun  
##   
## 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(lr\_tb)

## spec\_tbl\_df [100 x 5] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ phy : num [1:100] 43 54 60 57 60 42 48 46 57 59 ...  
## $ psy : num [1:100] 18 27 30 17 30 27 21 18 30 24 ...  
## $ cmmt: num [1:100] 28 28 26 23 29 26 23 28 23 26 ...  
## $ exp : Factor w/ 2 levels "No","Yes": 2 2 1 1 2 2 1 1 1 1 ...  
## $ chun: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 1 1 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. phy = col\_double(),  
## .. psy = col\_double(),  
## .. cmmt = col\_double(),  
## .. exp = col\_double(),  
## .. chun = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

lr\_tb

## # A tibble: 100 x 5  
## phy psy cmmt exp chun   
## <dbl> <dbl> <dbl> <fct> <fct>  
## 1 43 18 28 Yes No   
## 2 54 27 28 Yes No   
## 3 60 30 26 No No   
## 4 57 17 23 No No   
## 5 60 30 29 Yes No   
## 6 42 27 26 Yes Yes   
## 7 48 21 23 No No   
## 8 46 18 28 No No   
## 9 57 30 23 No No   
## 10 59 24 26 No No   
## # ... with 90 more rows

# 3.기본통계치 확인

skim(lr\_tb)

Data summary

|  |  |
| --- | --- |
| Name | lr\_tb |
| Number of rows | 100 |
| Number of columns | 5 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 2 |
| numeric | 3 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: factor**

| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| --- | --- | --- | --- | --- | --- |
| exp | 0 | 1 | FALSE | 2 | Yes: 52, No: 48 |
| chun | 0 | 1 | FALSE | 2 | No: 72, Yes: 28 |

**Variable type: numeric**

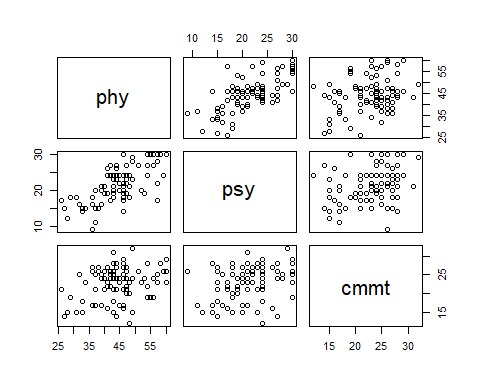
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| phy | 0 | 1 | 44.86 | 7.45 | 26 | 41 | 45 | 48 | 60 | ▁▂▇▃▃ |
| psy | 0 | 1 | 21.78 | 4.88 | 9 | 18 | 22 | 24 | 30 | ▁▅▇▇▆ |
| cmmt | 0 | 1 | 22.95 | 4.21 | 12 | 21 | 24 | 26 | 32 | ▂▃▇▇▁ |

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

## # A tibble: 3 x 13  
## variable n min max median q1 q3 iqr mad mean sd se  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 cmmt 100 12 32 24 21 26 5 4.45 23.0 4.21 0.421  
## 2 phy 100 26 60 45 41 48 7 4.45 44.9 7.45 0.745  
## 3 psy 100 9 30 22 18 24 6 4.45 21.8 4.88 0.488  
## # ... with 1 more variable: ci <dbl>

# 4.그래프 그리기

pairs( ~ phy+psy+cmmt, data=lr\_tb)



# 5.로지스틱 회귀분석

lr\_fit <- glm(chun ~ phy+psy+cmmt+exp,   
 family = binomial,   
 data=lr\_tb)

## ANOVA 분석

summary(lr\_fit)

##   
## Call:  
## glm(formula = chun ~ phy + psy + cmmt + exp, family = binomial,   
## data = lr\_tb)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.0764 -0.4233 -0.1697 0.2132 2.5153   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 14.80336 3.60443 4.107 4.01e-05 \*\*\*  
## phy -0.05610 0.06693 -0.838 0.4019   
## psy -0.06552 0.09872 -0.664 0.5069   
## cmmt -0.60321 0.12860 -4.691 2.72e-06 \*\*\*  
## expYes 2.02826 0.85186 2.381 0.0173 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 118.591 on 99 degrees of freedom  
## Residual deviance: 55.449 on 95 degrees of freedom  
## AIC: 65.449  
##   
## Number of Fisher Scoring iterations: 6

## 회귀계수

tidy(lr\_fit, conf.int = TRUE)

## # A tibble: 5 x 7  
## term estimate std.error statistic p.value conf.low conf.high  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 14.8 3.60 4.11 0.0000401 8.72 23.1   
## 2 phy -0.0561 0.0669 -0.838 0.402 -0.200 0.0680  
## 3 psy -0.0655 0.0987 -0.664 0.507 -0.267 0.126   
## 4 cmmt -0.603 0.129 -4.69 0.00000272 -0.897 -0.384   
## 5 expYes 2.03 0.852 2.38 0.0173 0.517 3.94

## 설명력R2

glance(lr\_fit)

## # A tibble: 1 x 8  
## null.deviance df.null logLik AIC BIC deviance df.residual nobs  
## <dbl> <int> <dbl> <dbl> <dbl> <dbl> <int> <int>  
## 1 119. 99 -27.7 65.4 78.5 55.4 95 100

## Odds 계산

tidy(lr\_fit, conf.int = TRUE) %>%  
 mutate(odds = exp(coef(lr\_fit))) %>%  
 write\_csv("table/lr\_tb.csv")