Simple linear regression-test

Youngjin\_Lee

2022 2 20

Table of Contents

통계분석 관계검정 : 상관분석(Simple linear regression) test 문제 일개 기업체에서 근무하고 있는 직원(100명)들의 정규적인 건강검진 결과의 일부 자료이다. 콜레스테롤이 높으면 중성지방도 높다고 말할 수 있는가? 그렇다면 콜레스테롤과 중성지방 사이의 관련성을 회귀식으로 추정하시오

# 1.기본 package 설정, library 로드

# 2.데이터 불러오기

slr\_tb <- read\_csv('data/SLR.csv',   
 col\_names = TRUE,  
 locale=locale('ko', encoding='euc-kr'), # 한글  
 na=".") %>%  
 mutate\_if(is.character, as.factor)

## Rows: 62 Columns: 2  
## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## dbl (2): col, fat  
##   
## 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(slr\_tb)

## spec\_tbl\_df [62 x 2] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ col: num [1:62] 146 190 179 164 185 ...  
## $ fat: num [1:62] 162 225 194 161 181 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. col = col\_double(),  
## .. fat = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

skim(slr\_tb)

Data summary

|  |  |
| --- | --- |
| Name | slr\_tb |
| Number of rows | 62 |
| Number of columns | 2 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| numeric | 2 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| col | 0 | 1 | 171.14 | 28.05 | 108.4 | 150.98 | 169.70 | 187.25 | 257.4 | ▂▇▇▃▁ |
| fat | 0 | 1 | 117.15 | 81.56 | 40.9 | 65.43 | 102.25 | 141.23 | 508.5 | ▇▃▁▁▁ |

# 3.기본통계치 확인

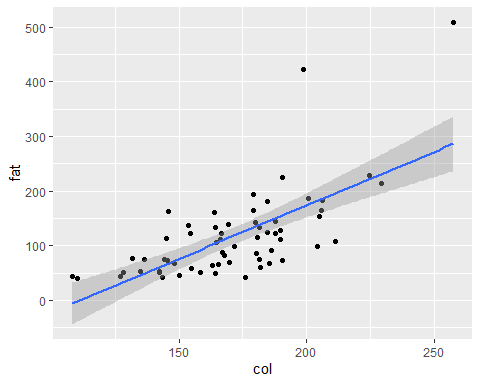
slr\_tb %>%  
 get\_summary\_stats(fat, col)

## # A tibble: 2 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 col 62 108. 257. 170. 151. 187. 36.3 26.8 171. 28.1 3.56  
## 2 fat 62 40.9 508. 102. 65.4 141. 75.8 56.5 117. 81.6 10.4   
## # ... with 1 more variable: ci <dbl>

# 4.그래프 그리기(산점도)

slr\_tb %>%  
 ggplot(mapping = aes(x = col,  
 y = fat)) +  
 geom\_jitter() +  
 geom\_smooth(method = "lm", se = TRUE)

## `geom\_smooth()` using formula 'y ~ x'



# 5.단순회귀분석

## 표준화 회귀계수

## 모델 생성

slr\_fit <- lm(fat ~ ., data = slr\_tb) %>%  
 lm.beta() # 표준화 회귀계수

## ANOVA 분석

anova(slr\_fit)

## Analysis of Variance Table  
##   
## Response: fat  
## Df Sum Sq Mean Sq F value Pr(>F)   
## col 1 184334 184334 49.938 1.925e-09 \*\*\*  
## Residuals 60 221476 3691   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## 회귀계수

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

## # A tibble: 2 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~ -218. 0 48.1 -4.54 2.78e-5 -96.2 96.2   
## 2 col 1.96 0.674 0.277 7.07 1.93e-9 0.119 1.23

## 설명력R2

glance(slr\_fit)

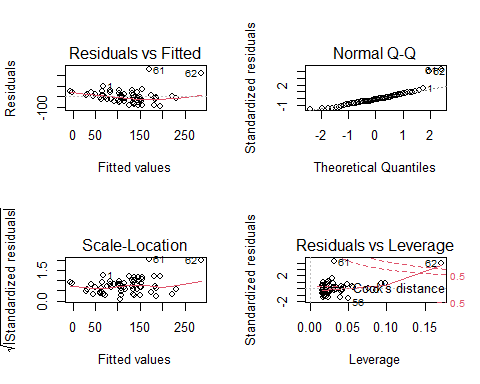
## Warning: Tidiers for objects of class lm.beta are not maintained by the broom  
## team, and are only supported through the lm tidier method. Please be cautious in  
## interpreting and reporting broom output.

## # 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 0.454 0.445 60.8 49.9 0.00000000193 1 -342. 689. 696.  
## # ... 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(slr\_fit)



par(opar)

## 잔차의 정규분포 검정

shapiro\_test(slr\_fit$residuals)

## # A tibble: 1 x 3  
## variable statistic p.value  
## <chr> <dbl> <dbl>  
## 1 slr\_fit$residuals 0.848 0.00000192

## 수치로 가정 검정(잔차의 등분산성 검정)

car::ncvTest(slr\_fit)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 33.2511, Df = 1, p = 8.0994e-09

## 이상치 검정, sd, hat, d 통합검정

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

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

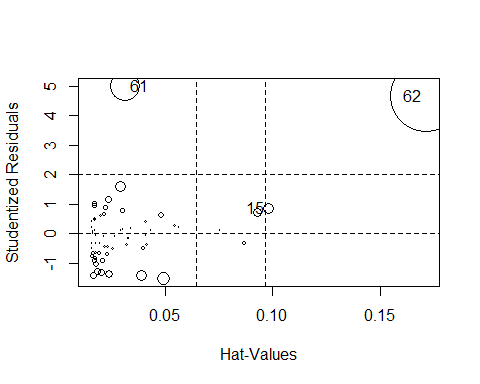
## Warning in plot.xy(xy, type, ...): "id.method"는 그래픽 매개변수가 아닙니다

## 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  
## 15 0.8628319 0.09811898 0.04067052  
## 61 4.9889620 0.03172057 0.29159034  
## 62 4.6626699 0.17111158 1.66756404

## 이상치 제거

slr\_tb <- slr\_tb[-c(61:62),]

# 7.단순회귀분석(이상치제거 다시분석)

## 모델 생성

slr\_fit <- lm(fat ~ ., data = slr\_tb) %>%  
 lm.beta() # 표준화 회귀계수

## ANOVA 분석

anova(slr\_fit)

## Analysis of Variance Table  
##   
## Response: fat  
## Df Sum Sq Mean Sq F value Pr(>F)   
## col 1 64716 64716 43.5 1.396e-08 \*\*\*  
## Residuals 58 86288 1488   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## 회귀계수

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

## # A tibble: 2 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~ -111. 0 33.2 -3.34 1.47e-3 -66.4 66.4   
## 2 col 1.28 0.655 0.194 6.60 1.40e-8 0.267 1.04

## 설명력R2

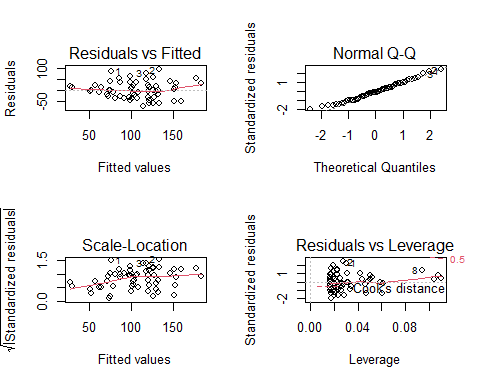
glance(slr\_fit)

## 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 0.429 0.419 38.6 43.5 0.0000000140 1 -303. 613. 619.  
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>

## 그림으로 가정 검정

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



par(opar)

## 잔차의 정규분포 검정

shapiro\_test(slr\_fit$residuals)

## # A tibble: 1 x 3  
## variable statistic p.value  
## <chr> <dbl> <dbl>  
## 1 slr\_fit$residuals 0.983 0.558

## 수치로 가정 검정

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

## 잔차의 등분산성 검정

car::ncvTest(slr\_fit)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 2.414986, Df = 1, p = 0.12018

## 이상치 검정, sd, hat, d 통합검정

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

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

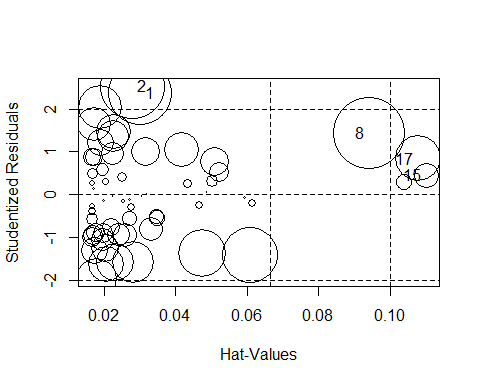
## Warning in plot.xy(xy, type, ...): "id.method"는 그래픽 매개변수가 아닙니다

## 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  
## 1 2.3656216 0.03031136 0.08104271  
## 2 2.5266123 0.02807170 0.08435904  
## 8 1.4232628 0.09430885 0.10363335  
## 15 0.4463291 0.11014609 0.01250169  
## 17 0.8372981 0.10802787 0.04267356

# 8.모델을 이용한 예측

## 콜레스테롤이 130, 150일 경우 예측값

slr\_tb\_new <- data.frame(col=c(130,150))  
predict(slr\_fit, newdata = slr\_tb\_new)

## 1 2   
## 55.35814 80.92389