regression model chapter 2

robinhwp

2023-03-14

# 제2장 분석사례(5강)

## 1) 자료읽기

library(xlsx)   
chemical=read.xlsx("./data/chemical.xlsx", 1)  
head(chemical)

## id speed temp loss  
## 1 1 70 20 15  
## 2 2 80 27 42  
## 3 3 75 25 37  
## 4 4 62 24 28  
## 5 5 65 23 20  
## 6 6 58 18 14

## 2) 기술통계량 및 상관계수 보기

# 자료 요약  
summary(chemical[,-1])

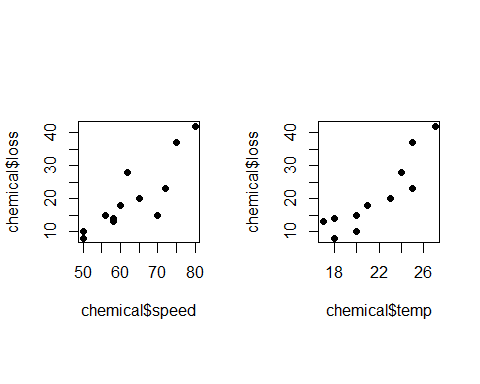
## speed temp loss   
## Min. :50.0 Min. :17.00 Min. : 8.00   
## 1st Qu.:57.5 1st Qu.:19.50 1st Qu.:13.75   
## Median :61.0 Median :20.50 Median :16.50   
## Mean :63.0 Mean :21.50 Mean :20.25   
## 3rd Qu.:70.5 3rd Qu.:24.25 3rd Qu.:24.25   
## Max. :80.0 Max. :27.00 Max. :42.00

# 상관계수  
cor(chemical[,-1])

## speed temp loss  
## speed 1.0000000 0.8023847 0.8548423  
## temp 0.8023847 1.0000000 0.8953498  
## loss 0.8548423 0.8953498 1.0000000

## 3) 산점도 그리기

par(mfrow=c(1,2), pty="s")   
plot(chemical$speed, chemical$loss, pch=19)  
plot(chemical$temp, chemical$loss, pch=19)



## 4) 회귀모형 적합하기

chemical.lm = lm(loss ~ speed+temp, data=chemical)   
summary(chemical.lm)

##   
## Call:  
## lm(formula = loss ~ speed + temp, data = chemical)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.7699 -2.4093 0.2795 3.4019 4.9654   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -47.6243 9.4580 -5.035 0.000704 \*\*\*  
## speed 0.4216 0.2350 1.794 0.106360   
## temp 1.9217 0.6977 2.754 0.022316 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.465 on 9 degrees of freedom  
## Multiple R-squared: 0.8539, Adjusted R-squared: 0.8214   
## F-statistic: 26.3 on 2 and 9 DF, p-value: 0.0001741

### 추정된 회귀방정식 1)

# 추정된 회귀방정식 1) (chemical.lm$model과 coef(chemical.lm) 함수를 이용)  
str\_l = paste0("hat\_", names(chemical.lm$model)[1])  
str\_r = paste0( round(coef(chemical.lm)["(Intercept)"], 3))  
for(i in 2:length(coef(chemical.lm)))   
{  
 str\_r = paste0( str\_r, " + ", round(coef(chemical.lm)[names(chemical.lm$model)[i]], 3),  
 "\*", names(chemical.lm$model)[i])  
}  
paste0("추정된 회귀방정식: ", str\_l, " = ", str\_r)

## [1] "추정된 회귀방정식: hat\_loss = -47.624 + 0.422\*speed + 1.922\*temp"

### 추정된 회귀방정식 2)

# 추정된 회귀방정식 2) 최소제곱법으로 β를 행렬방정식으로 계산  
mX=as.matrix(cbind(1, chemical$speed, chemical$temp))  
mY=as.matrix(chemical$loss)  
mβ=solve(t(mX)%\*%mX)%\*%t(mX)%\*%mY  
paste0(" 최소제곱법 이용 : hat\_loss = ", round(mβ, 3)[1], " + ",  
 round(mβ, 3)[2], "\*speed + ",  
 round(mβ, 3)[3], "\*temp")

## [1] " 최소제곱법 이용 : hat\_loss = -47.624 + 0.422\*speed + 1.922\*temp"

### 결정계수

# 결정계수  
chemical.lm.summary=summary(chemical.lm)  
paste0("결정계수 R.squared = ", round(chemical.lm.summary$r.squared, 3), "으로 ",  
 round(chemical.lm.summary$r.squared\*100,1), "% 설설명력이 있다.")

## [1] "결정계수 R.squared = 0.854으로 85.4% 설설명력이 있다."

### p-value

# p-value  
α = 0.05  
for(i in 2:length(coef(chemical.lm)))   
{  
 p.value = round(chemical.lm.summary$coefficients[names(chemical.lm$model)[i], "Pr(>|t|)"], 5)  
 if (p.value < α )  
 {  
 print(paste0( names(chemical.lm$model)[i], "의 p-value 가 ", p.value, "으로서 ",   
 names(chemical.lm$model)[1], "를(을) 설명하는데 유의하다"))  
 }  
 else  
 {  
 print(paste0( names(chemical.lm$model)[i], "의 p-value 가 ", p.value, "으로서 ",   
 names(chemical.lm$model)[1], "를(을) 설명하는데 그리 큰 영향을 준다고 할 수 없다."))  
 }  
}

## [1] "speed의 p-value 가 0.10636으로서 loss를(을) 설명하는데 그리 큰 영향을 준다고 할 수 없다."  
## [1] "temp의 p-value 가 0.02232으로서 loss를(을) 설명하는데 유의하다"

## 5) 분산분석표 구하기

# 4) 분산분석표 구하기  
chemical.lm.anova=anova(chemical.lm)  
AVT = matrix(c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0), nrow=3)  
colnames(AVT)=colnames(chemical.lm.anova)  
rownames(AVT)=c("Regression", "Residuals", "Total")  
  
for(i in 1:(length(rownames(AVT))-1))  
{  
 AVT["Regression", "Df"]=AVT["Regression", "Df"]+as.double(chemical.lm.anova[i,"Df"])  
 AVT["Regression", "Sum Sq"]=AVT["Regression", "Sum Sq"]+as.double(chemical.lm.anova[i,"Sum Sq"])  
}  
AVT["Residuals", "Df"]=as.double(chemical.lm.anova["Residuals","Df"])  
AVT["Residuals", "Sum Sq"]=as.double(chemical.lm.anova["Residuals","Sum Sq"])  
AVT["Total", "Df"]=AVT["Regression", "Df"]+AVT["Residuals", "Df"]  
AVT["Total", "Sum Sq"]=AVT["Regression", "Sum Sq"]+AVT["Residuals", "Sum Sq"]  
AVT["Regression", "Mean Sq"] = AVT["Regression", "Sum Sq"] / AVT["Regression", "Df"]  
AVT["Residuals", "Mean Sq"] = AVT["Residuals", "Sum Sq"] / AVT["Residuals", "Df"]  
AVT["Regression", "F value"] = round(AVT["Regression", "Mean Sq"] / AVT["Residuals", "Mean Sq"], 1)  
AVT["Regression", "Pr(>F)"] = round(1-pf(AVT["Regression", "F value"],   
 AVT["Regression", "Df"],   
 AVT["Residuals", "Df"]), 6)  
print(AVT)

## Df Sum Sq Mean Sq F value Pr(>F)  
## Regression 2 1048.8114 524.40571 26.3 0.000174  
## Residuals 9 179.4386 19.93762 0.0 0.000000  
## Total 11 1228.2500 0.00000 0.0 0.000000

## 6) 잔차 산점도 : (독립변수, 잔차)

# 6) 잔차 산점도 : (독립변수, 잔차)  
par(mfrow=c(1,2), pty="s")  
plot(chemical$speed, chemical.lm$resid, pch=19)  
abline(h=0, lty=2)  
for (i in 1:length(chemical$speed))  
{  
 if(abs(chemical.lm$resid[i]) > 3.7)  
 {  
 text(chemical$speed[i]+0.3, chemical.lm$resid[i]-0.4, as.character(i))  
 }  
}  
  
plot(chemical$temp, chemical.lm$resid, pch=19)  
abline(h=0, lty=2)  
for (i in 1:length(chemical$temp))  
{  
 if(abs(chemical.lm$resid[i]) > 3.7)  
 {  
 text(chemical$temp[i]+0.3, chemical.lm$resid[i]-0.3, as.character(i))   
 }  
}

