R기초_이근백교수님 3일차 연습문제 풀이

#연습문제 p227

5. 연습문제

- 사원들의 발표능력, 분석능력 및 응용력이 사원들의 직무성적에 어떠한 영향을 미치는지를 분석하라.

사원	직무성적	발표능력	분석능력	응용력
1	87	9.2	8.4	8.7
2	93	9.4	9.3	9.4
3	91	9.5	9.2	9.6
4	85	8.7	7.9	8.9
5	86	8.8	8.1	8.6
6	97	9.6	9.8	9.3
7	90	9.2	9.0	9.0
8	93	9.5	9.4	9.2
9	88	8.5	8.6	8.9
10	96	9.6	9.7	9.5
11	86	9.4	8.3	8.8
12	89	8.7	8.7	9.0
13	94	9.6	9.2	9.1
14	91	9.2	9.0	9.5
15	95	9.7	9.3	9.1

<코드>

```
#데이터 작성

work <- c(87, 93, 91, 85, 86, 97, 90, 93, 88, 96, 86, 89, 94, 91, 95)

ppt<-c(9.2, 9.4, 9.5, 8.7, 8.8, 9.6, 9.2, 9.5, 8.5, 9.6, 9.4, 8.7, 9.6,
9.2, 9.7)

alz <-c(8.4, 9.3, 9.2, 7.9, 8.1, 9.8, 9.0, 9.4, 8.6, 9.7, 8.3, 8.7, 9.2,
9.0, 9.3)

apply <- c(8.7, 9.4, 9.6, 8.9, 8.6, 9.3, 9.0, 9.2, 8.9, 9.5, 8.8, 9.0,
9.1, 9.5, 9.1)

#draw a plot

plot(work, ppt)

plot(work, alz)
```

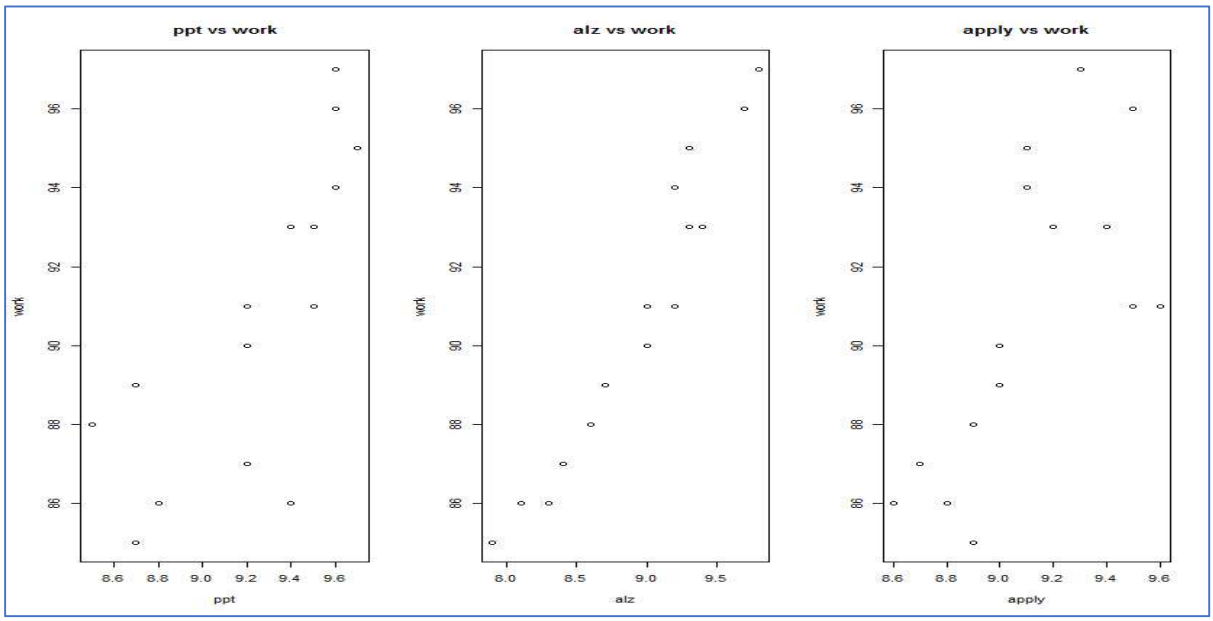
```
plot(work, apply)

fit1 <- lm(work ~ ppt+alz+apply)

summary(fit1)
```

<문제 해설>

1) Work, ppt, alz, apply 오브젝트에 문제에서 볼 수 있는 데이터를 입력합니다.

2) 문제를 보니 work가 반응변수이고, 나머지는 설명변수임을 알 수 있습니다.

3) 설명변수와 반응변수 사이에 선형관계가 있는지 plot을 그려서 확인해봅니다. 결과는 아래와 같고, 어느정도 선형관계가 있음을 알 수 있습니다..



<plot>

4) 이제 lm function을 이용해서 다중선형회귀분석을 실시합니다. 그 결과를 fit1 오브젝트에 넣고 summary함수를 통해서 그 결과를 확인합니다. 결과는 아래와 같습니다.

```
Call:
lm(formula = work ~ ppt + alz + apply)

Residuals:
     Min       1Q   Median       3Q      Max
-1.3471 -0.8429  0.0339  0.4225  1.4173

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  36.4740    10.2903   3.545   0.0046 **
ppt           0.6891     1.0399   0.663   0.5212
alz           6.7586     0.9430   7.167 1.83e-05 ***
apply        -1.3660     1.4095  -0.969   0.3533
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.01 on 11 degrees of freedom
Multiple R-squared:  0.9463,    Adjusted R-squared:  0.9317
F-statistic: 64.62 on 3 and 11 DF,  p-value: 2.853e-07
```

5) 결과창 해석은 다음과 같습니다.

설명변수들의 coefficient 의 estimate 값을 봄으로 각 변수들의 영향력을 판단할 수 있습니다. 예를 들면, 다른 변수들이 고정되어있을 때, alz value 가 1 단위 증가한다면 work value 는 6.7 만큼 증가한다고 이야기할 수 있습니다. 그러나 지금 이 결과창에서 p-value 값을 보시면 ppt, apply 는 통계적으로 유의하지 않음을 알 수 있습니다. 이러한 결과가 나타나는 이유들 중 가장 주된 이유는 변수들간의 상관관계(다중공선성)가 있을 가능성이 높습니다.

#연습문제 p274

4. 연습문제

12명의 평가요원이 4종류의 오렌지 A, B, C, D 에 신 맛 차이를 평가한다. 다음 자료는 평가요원이 오렌지마다 12명씩 랜덤 배치되어 9점 척도를 사용하여 평가한 결과이다. 오렌지 데이터에 대하여 신 맛에 차이가 있는지 검정하고, 어떤 오렌지 간에 차이가 있는지 튜키 검정을 통해 분석하세요.

A오렌지	B오렌지	C오렌지	D오렌지
8	6	8	9
8	5	8	4
7	7	8	6
9	7	5	5
8	6	6	4
7	7	7	4
9	6	6	5
8	8	7	6
6	7	7	7
8	6	6	5
7	8	7	8
6	7	8	4

<코드>

```
yy=c(8,8,7,9,8,7,9,8,6,8,7,6,6,5,7,7,6,7,6,8,7,6,8,7,8,8,8,5,6,7,6,7,7,6,7
,8,9,4,6,5,4,4,5,6,7,5,8,4)

xx=c(rep("a",12),rep("b",12),rep("c",12),rep("d",12))

dfr <- data.frame(yy,xx)

anova <-aov(yy~xx, data=dfr)

summary(anova)
```

```
TukeyHSD(anova)
```

<문제 해설>

1) 먼저 문제에 나와있는 데이터를 yy, xx 오브젝트에 각각 입력합니다.

2) Data.frame 함수를 이용해서 yy 와 xx 로 data.frame 을 만듭니다.

3) 이제 aov 함수를 이용해서 분산분석을 이용하고 summary 함수를 이용해서 그 결과를 봅니다.

```
> anova <-aov(yy~xx, data=dfr)
> summary(anova)
             Df Sum Sq Mean Sq F value Pr(>F)
xx            3  24.90   8.299   5.945 0.0017 **
Residuals    44  61.42   1.396
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4) 위 결과를 보면, p-value 값이 주어진 유의수준보다 훨씬 작기 때문에, 귀무가설, 즉 그룹간의 평균차이가 없다,를 기각하게 됩니다.

5) TukeyHsd 함수의 기능은 다음과 같습니다.

TukeyHSD {stats} R Documentation

Compute Tukey Honest Significant Differences

Description

Create a set of confidence intervals on the differences between the means of the levels of a factor with the specified family-wise probability of coverage. The intervals are based on the Studentized range statistic, Tukey's 'Honest Significant Difference' method.

Usage

```
TukeyHSD(x, which, ordered = FALSE, conf.level = 0.95, ...)
```

즉, 요인 내 수준간의 평균의 차이에 대한 신뢰구간을 제공해주는 함수입니다.

6) 이 함수를 이용한 결과는 아래와 같습니다.

```
$xx
          diff       lwr          upr      p adj
b-a -0.9166667 -2.204481  0.37114792 0.2426562
c-a -0.6666667 -1.954481  0.62114792 0.5169920
d-a -2.0000000 -3.287815 -0.71218541 0.0008448
c-b  0.2500000 -1.037815  1.53781459 0.9542572
d-b -1.0833333 -2.371148  0.20448125 0.1268467
d-c -1.3333333 -2.621148 -0.04551875 0.0399184
```

신뢰구간이 0 을 포함하지 않거나 p-value 값이 유의수준 0.05 보다 작은 값을 찾게되면 d-a 와 d-c 의 간의 변수가 유의한 것으로 보입니다. 따라서 a 오렌지와 d 오렌지, c 오렌지와 d 오렌지 간에는 신 맛에 차이가 있다고 할 수 있습니다.