M83330 예측분석 2주차 과제

1번 문제

1) 답: 0.5313237

```
# 엑셀 클립보드 가져오기
mydf <- read.table(pipe("pbpaste"), header = T, sep="₩t")

# 상관계수 구하기
cor(mydf$Last.Year, mydf$This.Year)
```

```
[1] 0.5313237
```

```
# 상관계수 유의수준 구하기
cor.test(mydf$Last.Year, mydf$This.Year)
```

```
> cor.test(mydf$Last.Year, mydf$This.Year)
Pearson's product-moment correlation

data:  mydf$Last.Year and mydf$This.Year
t = 2.429, df = 15, p-value = 0.02818
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.06805998 0.80610662
sample estimates:
      cor
0.5313237
```

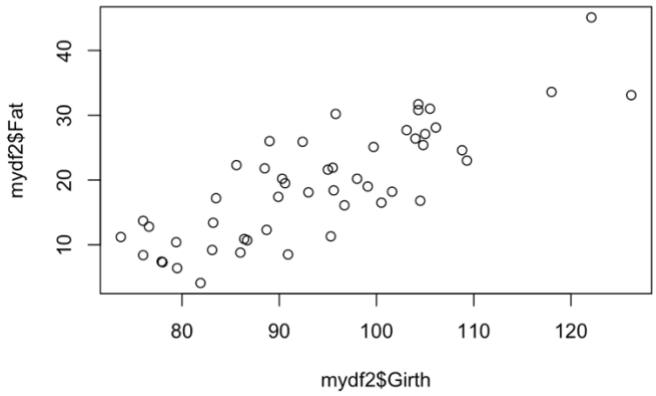
2) 답: 0.028, 즉 유의 수준 약 2~3% 수준에서 유의하다 볼 수 있으므로 1% 유의수준에서는 H0를 기각하지 못한다. 따라서, 1% 유의수준에서는 상관계수 r이 통계적으로 유의하지 못하다.

2번 문제

산포도 그리기

plot(formula=mydf2$Fat~mydf2$Girth)

1)



상관계수 구하기

cor(mydf2$Girth, mydf2$Fat)

```
[1] 0.8188484
```

모델 생성 후 summary

modelFat = lm(mydf2$Fat~mydf2$Girth, data= mydf2)

summary(modelFat)

```
> summary(modelFat)
Call:
lm(formula = mydf2$Fat ~ mydf2$Girth, data = mydf2)

Residuals:
     Min       1Q   Median       3Q      Max
-8.9392 -3.8714 -0.0414  3.6328  9.8672

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept) -36.23970       5.66898    -6.393 6.28e-08 ***
mydf2$Girth     0.59053       0.05975     9.883 3.71e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 5.086 on 48 degrees of freedom
Multiple R-squared:  0.6705,    Adjusted R-squared:  0.6636
F-statistic: 97.68 on 1 and 48 DF,  p-value: 3.714e-13
```

2) 답: 상관계수는 0.81 이고 결정계수는 0.67
3) 답: 단순회귀분석일 경우에는 상관계수의 제곱이 결정계수가 됨

```
# ols로 절편과 기울기 구하기
from statsmodels.formula.api import ols
res = ols('Fat ~ Girth', data=mydf2).fit()
res.summary()
```

4) 답: 기울기:0.5905, 절편: -36.2397
5) 답: Girth 가 1단위 증가함에 따라 Fat이 약 0.59 증가하는 것을 의미함
6) 답: H0(귀무가설) =기울기는 0이다 / H1(대립가설) = 기울기는 0이 아니다
7) 답: p-value가 3.71e-13이므로, 유의수준 5%보다 작기 때문에 귀무가설을 기각하므로 기울기는 0이 아니다. 이 모델은 유의하다 볼 수 있다.
8) 답: H0(귀무가설): 등분산성이 있다(집단간 분산이 같다) / H1(대립가설): 등분산성이 없다(집단간 분산이 다르다.)

```
# anova 그리기
g = lm(mydf2$Fat~mydf2$Girth)
anova(g)
```

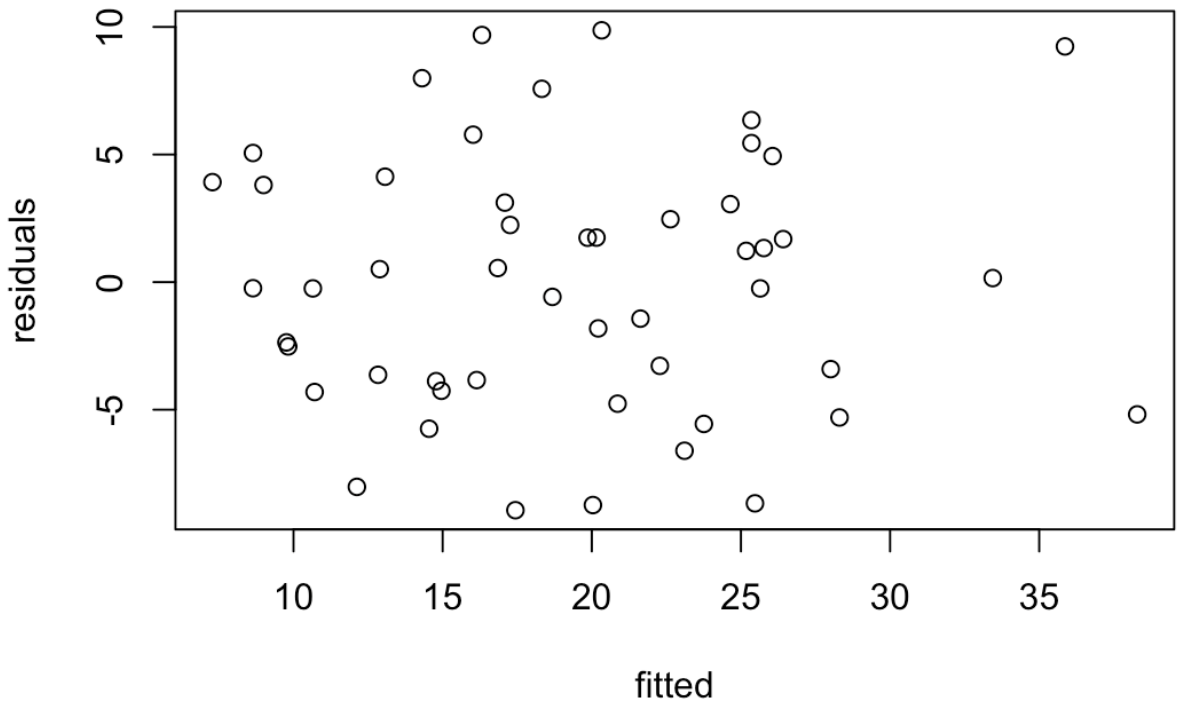
```
> anova(g);
Analysis of Variance Table

Response: mydf2$Fat
              Df Sum Sq Mean Sq F value      Pr(>F)
mydf2$Girth    1 2527.1 2527.12  97.681 3.714e-13 ***
Residuals     48 1241.8    25.87
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

9) 답: f분포에 대한 t-value가 3.714e-13(***) 이므로, 유의수준 5%보다 작기 때문에 귀무가설을 기각하지 않는다. 즉, 해당 모델은 등분산성이 있다.

```
# 잔차분석
```

g = lm(mydf2$Fat~mydf2$Girth)
plot(fitted(g), residuals(g),xlab="fitted",ylab="residuals")

10)



leverage가 3/n 보다 크면 이상치, 즉 0.06보다 크면 이상치이다.
# leverate point로 이상치 분석
ginf=influence(g)
ginf$hat

```
> ginf$hat[ginf$hat>0.06]
         2          5         10         28         45         48
0.06526755 0.09874332 0.06231805 0.06526755 0.16208213 0.07749436
        50
0.12809308
```

11) 답: 0.06보다 큰 2,5,10,28,45,48,50이 이상치이다.

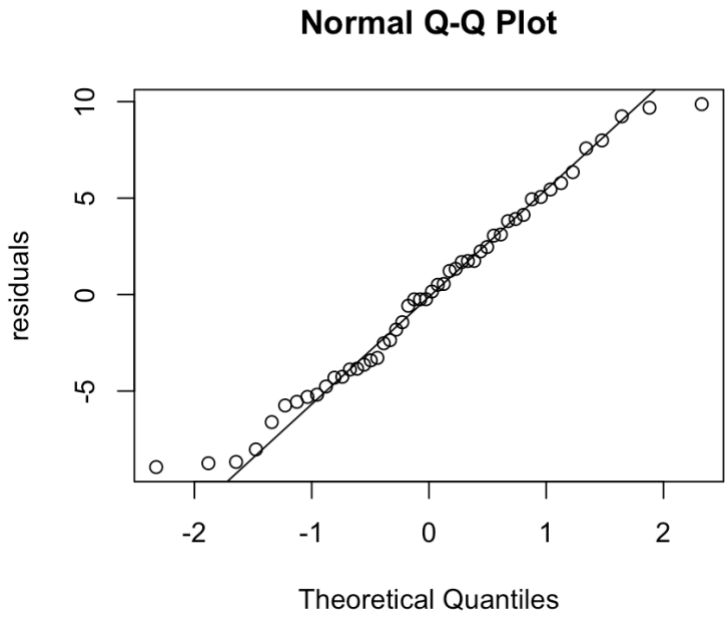
studentized deleted residuals의 절대값이 2 이상이면 이상치
# studentized deleted residuals로 이상치 분석
r1=rstudent(g)
abs(r1)[abs(r1)>2]

```
> abs(r1)[abs(r1)>2]
      37       50
2.022098 2.004957
```

12) 답: 37, 50

13) 답: 없음

14)




```
# 잔차의 정규분포 검증
ks.test(residuals(g), "pnorm", m=mean(residuals(g)), sd=sd(residuals(g))); # K-S test
```

```
data:  residuals(g)
D = 0.082751, p-value = 0.8553
alternative hypothesis: two-sided
```

15) 답: p-value는 0.8553이므로 유의수준 10%(0.1)에서 귀무가설을 기각할 수 없다. 즉, 잔차는 정규분포를 따른다.

```
# Breusch-Pagan test
bptest(Fat~Girth, data=mydf2)
```

```
studentized Breusch-Pagan test
data:  Fat ~ Girth
BP = 0.64435, df = 1, p-value = 0.4221
```

16) 답: p-value = 0.4221이므로 유의수준 10%(0.1)에서 귀무가설을 기각할 수 없다. 즉, 오차는 동분산이다.

```
# Durbin-Watson test
dwtest(Fat~Girth, data=mydf2)
```

```
Durbin-Watson test
data:  Fat ~ Girth
DW = 2.4016, p-value = 0.9247
alternative hypothesis: true autocorrelation is greater than 0
```

17) 답: p-value = 0.9247이므로 유의수준 10%(0.1)에서 귀무가설을 기각할 수 없다. 즉, 잔차의 상관성이 존재하지 않는다.