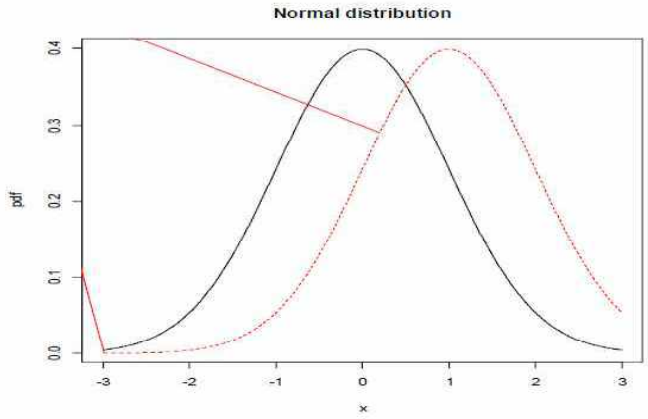
R기초_이근백교수님 2일차 연습문제 풀이

#연습문제 p105

5. 연습문제

- 검정선은 평균이 0, 분산이 1인 표준정규분포의 확률밀도함수를, 빨간 선은 평균이 1, 분산이 1인 표준정규분포의 확률밀도함수를 그린 것이다. R 함수 'dnorm'을 이용하여 다음 그림을 작성해라.



2. R 그래픽스 105

<코드>

```
x <-seq(-3,3,length.out=1000)
plot(x, dnorm(x, mean=0, sd=1), type='l',xlim=c(-3,3), main="Normal distribution, X~N(0,1)")
lines(x,dnorm(x, mean=1, sd=1), col="2", xlim=c(-3,3))
```

<해설>

1) Seq(-3,3, length.out=1000) 함수를 통해서 -3에서 3사이 실수들을 동일한 간격으로 1000개를 생성합니다. Ex) seq(0,2, length.out=3)을 작성하시면 0,1,2라는 값을 return합니다. 위 코드의 결과를 x라는 오브젝트에 저장합니다.

2) 이제 이 x라는 오브젝트를 활용하여 plot을 그릴 것입니다. plot함수를 사용하도록 합니다. plot(x, dnorm(x, mean=0, sd=1), type='l',xlim=c(-3,3), main="Normal distribution, X~N(0,1)") 함수에 대해서 설명드리면 다음과 같습니다. 먼저 dnorm(x, mean=0, sd=1)함수에 대해서 알아보겠습니다. ?dnorm 을 입력함으로 dnorm함수의 기능을 확인할 수 있습니다. Dnorm(x, mean=0, sd=1)은 평균이 0, 표준편차가 1일 때, x 벡터 속에

입력되어있는 각 value들의 정규분포 density를 return해줍니다. Plot(x, dnorm(x, mean=0, sd=1)이라는 함수는 x축에 x오브젝트를, y축에는 dnorm(x,mean=0,sd=1)을 입력하여 plot을 그리는 함수입니다. 그 뒤의 argument 중 type은 plot의 타입(점, 선 등등)을 의미합니다. 'l', 'p', 'b' 등이 있으며 ?plot을 통해서 확인 가능합니다. Xlim argument는 엑스 축의 정의역 range를 의미합니다. Main argument는 그래프의 overall title을 의미합니다.

3) 위 코드를 통해서 평균이 0, 표준편차가 1인 정규분포 plot을 그렸습니다. 이제 문제에서 요구하는 것처럼 이 plot 위에 평균이 1, 표준편차가 1인 정규분포를 겹쳐서 그리고 싶습니다. 이 때 사용할 수 있는 함수가 lines 함수입니다. ?lines를 통해서 lines 함수에 대한 설명을 볼 수 있습니다. lines(x,dnorm(x, mean=1, sd=1), col="2", xlim=c(-3,3))는 x축에는 x오브젝트 속의 value를, y축에는 dnorm(x, mean=1, sd=1)의 value를 토대로 line을 그리며, line의 color는 2번에 해당하는 색으로 함을 의미합니다.

4) 위 코드를 통해서 원하는 plot과 line이 구현되었음을 확인하였습니다.

#연습문제 p174

5. 연습문제

- 일반적으로 중학교 1학년 학생들의 평균 신장은 150cm정도 되는 것으로 받아들여진다. 이에 따라 어느 중학교 1학년 학생들의 평균 신장이 150cm라고 말할 수 있는지를 조사하기 위하여 1학년 학생들 30명의 키를 측정한 결과가 다음과 같다. 유의수준 5%에서 이 학교 1학년 학생들의 평균 신장이 150cm라고 할 수 있는지 검정하라.

148 150 149 144 152 150 155 147 148 151 150 149 150 144 147
150 153 147 152 150 151 149 149 153 147 152 160 165 140 141

- 회사는 자체 개발한 교육 프로그램이 효과가 있는지 여부를 분석하기로 하였다. 교육 프로그램에 참가한 10명의 성적이 다음과 같다고 하였을 때, 이 프로그램은 효과가 있다고 할 수 있는가?

	1	2	3	4	5	6	7	8	9	10
교육 후	70	62	54	82	75	64	58	57	80	63
교육 전	68	62	50	75	76	57	60	53	74	60

<1번문제 코드>

```
install.packages("TeachingDemos")
#z.test를 위한 패키지 설치
library(TeachingDemos)
#패키지 load

a<-
c(148,150,149,144,152,150,155,147,148,151,150,149,150,144,147,150,153,147,152,150,151,149,149,153,147,152,160,165,140,141)
Mu0 <-150
sd0 <-sd(a)
#30명 -> 대표본, 모집단의 분포에 관계없이 z검정가능.

z.test(a, mu=150,sd=sd0)

# p-value값을 봤을 때, 귀무가설을 기각할 수 없다.
```

<1번문제 해설>

1) 먼저 z-test를 실행하기위해서 패키지를 설치하고, 이 패키지를 load 시켜주겠습니다. install.packages("TeachingDemos")를 통해서 패키지 설치가 가능하고, library함수를 통해서 load할 수 있습니다.

2) 이제 문제에서 제공한 표본들을 a라는 오브젝트에 저장하겠습니다.

3) 그리고 Mu0 라는 오브젝트에 문제에서 제시한 모집단 평균 150을 저장합니다.

4) Sd0 라는 오브젝트에 sd(a)를 통해서 나온 결과값을 저장합니다. Sd(a)는 a 오브젝트 속에 속한 value들의 표준편차를 return시켜주는 함수입니다.

5) 총 표본은 30개입니다. 이는 손으로 세어보셔도 되지만, length(a)를 통해서도 확인할 수 있습니다.

6) 표본갯수가 충분히 많기 때문에, 모집단의 분포에 관계없이 z-test를 사용할 수 있습니다. 모집단의 표준편차를 안다면 모집단의 표준편차를 사용하면 되지만, 모집단의 표준편차를 모른다면 표본집단의 표준편차를 사용해도 무방합니다.

7) Ztest(a, mu=150, sd=sd0)를 작성함으로 z-test를 실시합니다. 결과는 아래 그림과 같습니다.

```
> z.test(a, mu=150,sd=sd0)

        One Sample z-test

data:  a
z = -0.26366, n = 30.00000, Std. Dev. = 4.84720, Std. Dev. of the sample mean = 0.88497, p-value = 0.792
alternative hypothesis: true mean is not equal to 150
95 percent confidence interval:
 148.0321 151.5012
sample estimates:
mean of a
 149.7667
```

<z.test(a, mu=150, sd=sd0) 결과>

8) 위 결과를 해석하자면 다음과 같습니다. 먼저 p-value를 봤을 때, 이는 문제에서 설정한 유의수준 5%(0.05)보다 큰 값입니다. 따라서 귀무가설, 즉 모집단의 평균은 150이다, 라는 가설을 5% 유의수준 하에서 기각할 수 없습니다.

<2번문제 코드>

```
before <-c(70,62,54,82,75,64,58,57,80,63)
after <-c(68,62,50,75,76,57,60,53,74,60)

#paired-t-test를보자.
t.test(before,after, mu=0, paired=TRUE)

#p-value를 보니 0.01661, 알파값보다 작으므로 귀무가설을 기각한다.
#즉, 두집단의 평균이 같다는 귀무가설을 기각한다.
```

<2번문제 해설>

2번문제를 보니, 같은 대상에 대해서 교육 전과 교육 후의 효과의 차이가 존재하는지 통계적으로 검증하고 싶어함을 알 수 있습니다. 이는 대응표본분석에 해당합니다.

1) before라는 오브젝트에 교육 전에 대한 데이터를 저장하고, after라는 오브젝트에 교육 후에 대한 데이터를 저장합니다.

2) 표본숫자가 충분하지 않다고 판단되기 때문에 paired-t test를 사용하겠습니다. t.test(before, after, mu=0, paired=TRUE) 코드를 설명드리자면 다음과 같습니다. Before

object에 속한 값과 after object에 속한 값의 모평균차가 0인지 검증하고자 하며, 여기서는 이표본분석이 아닌 대응표본분석이므로 paired=TRUE 옵션을 적어줍니다. 결과는 아래와 같습니다.

```
> before <-c(70,62,54,82,75,64,58,57,80,63)
> after <-c(68,62,50,75,76,57,60,53,74,60)
>
> #paired-t-test를보자.
> t.test(before,after, mu=0, paired=TRUE)

        Paired t-test

data:  before and after
t = 2.9355, df = 9, p-value = 0.01661
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.6881192 5.3118808
sample estimates:
mean of the differences
                      3
```

< t.test(before, after, mu=0, paired=TRUE) 결과>

3) 위 결과를 해석하면 다음과 같습니다. 앞선 문제와 마찬가지로 유의수준 5% 하에서 검정하고 싶다고 가정하겠습니다. P-value 값을 보니 유의수준 5%(0.05)보다 작습니다. 따라서 귀무가설, 즉 교육 전과 교육 후의 효과의 차이가 없다, 은 5% 유의수준 하에서 기각됩니다.

#연습문제 p175

5. 연습문제

- 비타민 C가 감기예방에 효과가 있는지 알아보기 위해 200명의 지원자를 랜덤하게 두 그룹으로 나누어 100명에게는 위약을 투여하고, 100명에게는 매일 1g의 비타민 를 투여하였다. 한 겨울이 지난 후에 감기에 걸렸던 사람의 수를 조사한 결과가 다음과 같았다.

	감기 걸림	감기 안걸림	계
위약	40	60	100
비타민 C	20	80	100

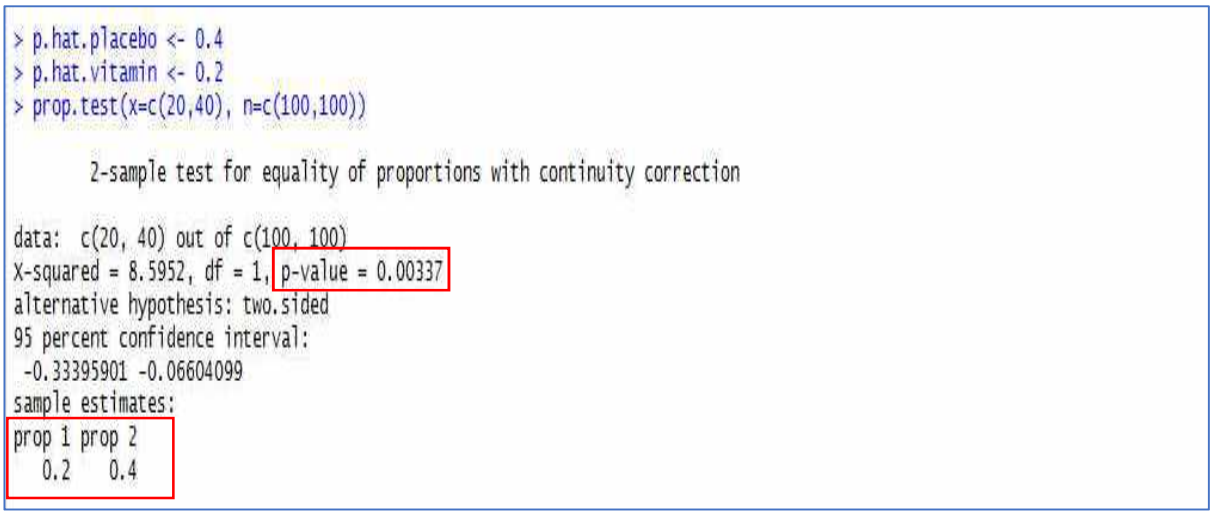
➢ 위약 그룹과 비타민 C 그룹의 감기 발병률을 계산하여라.

➢ 비타민 C가 감기 예방에 효과가 있다고 할 수 있는가?

<1번문제 코드>

```
#위약그룹의 발병률(표본) = 40/100 =0.4
#비타민C그룹의 발병률(표본) = 20/100 =0.2

#이 문제에서는 모비율의 차를 비교해보고 싶어함.
p.hat.placebo <- 0.4
p.hat.vitamin <- 0.2
prop.test(x=c(20,40), n=c(100,100))
#위 문제들과 마찬가지로 p-value값을 보니 두 집단간의 차이가 있다고 말할 수 있다.
```

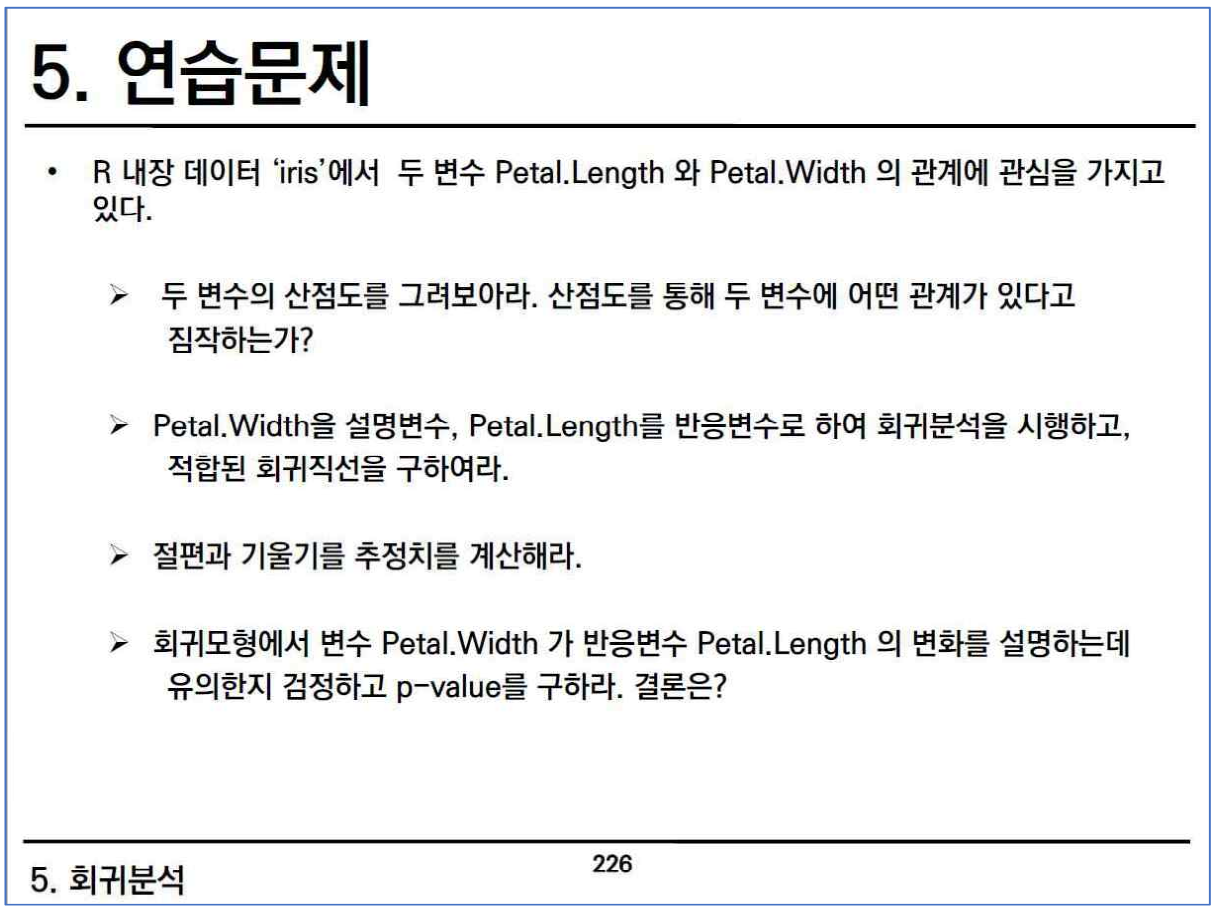


< prop.test(x=c(20,40), n=c(100,100)) 결과>

<문제 해설>

위 결과창을 해석하면 다음과 같습니다. prop.test(x=c(20,40), n=c(100,100))을 먼저 설명하자면, prop.test 의 argument 로 x 와 n 을 입력해야 합니다. X 는 성공횟수를 n 은 시행횟수를 넣어줍니다. 그 결과 p-value 값이 유의수준 5%보다 작으므로, 귀무가설, placebe 효과와 vitamin 효과가 차이가 없다, 을 기각합니다.

#연습문제 p226



5. 연습문제

- R 내장 데이터 'iris'에서 두 변수 Petal.Length 와 Petal.Width 의 관계에 관심을 가지고 있다.
  - ➢ 두 변수의 산점도를 그려보아라. 산점도를 통해 두 변수에 어떤 관계가 있다고 짐작하는가?
  - ➢ Petal.Width을 설명변수, Petal.Length를 반응변수로 하여 회귀분석을 시행하고, 적합된 회귀직선을 구하여라.
  - ➢ 절편과 기울기를 추정치를 계산해라.
  - ➢ 회귀모형에서 변수 Petal.Width 가 반응변수 Petal.Length 의 변화를 설명하는데 유의한지 검정하고 p-value를 구하라. 결론은?

5. 회귀분석 226

<코드>

```
attach(iris)
plot(Petal.Width, Petal.Length)
#산점도를 그려보았다.
result <- lm(Petal.Length~Petal.Width)
abline(result$coefficients)
#회귀선을 그려보았다.
#이를 통해서 선형관계가 있음을 볼 수 있다.

lm(Petal.Length~Petal.Width)
#결과값을 보면 B0는 1.084, B1은 2.230임을 알 수 있다.
#따라서 회귀직선은 y = 1.084+ 2.230x이며, 여기서의 x는 Petal.Width, Y는 Petal.Length이다.
#이를 통해서 절편의 추정치는 1.084이고 기울기의 추정치는 2.230임을 알 수 있다.
summary(result)
#이를 통해서 Petal.Width의 P-value값이 무척 작음을 알 수 있다.
```

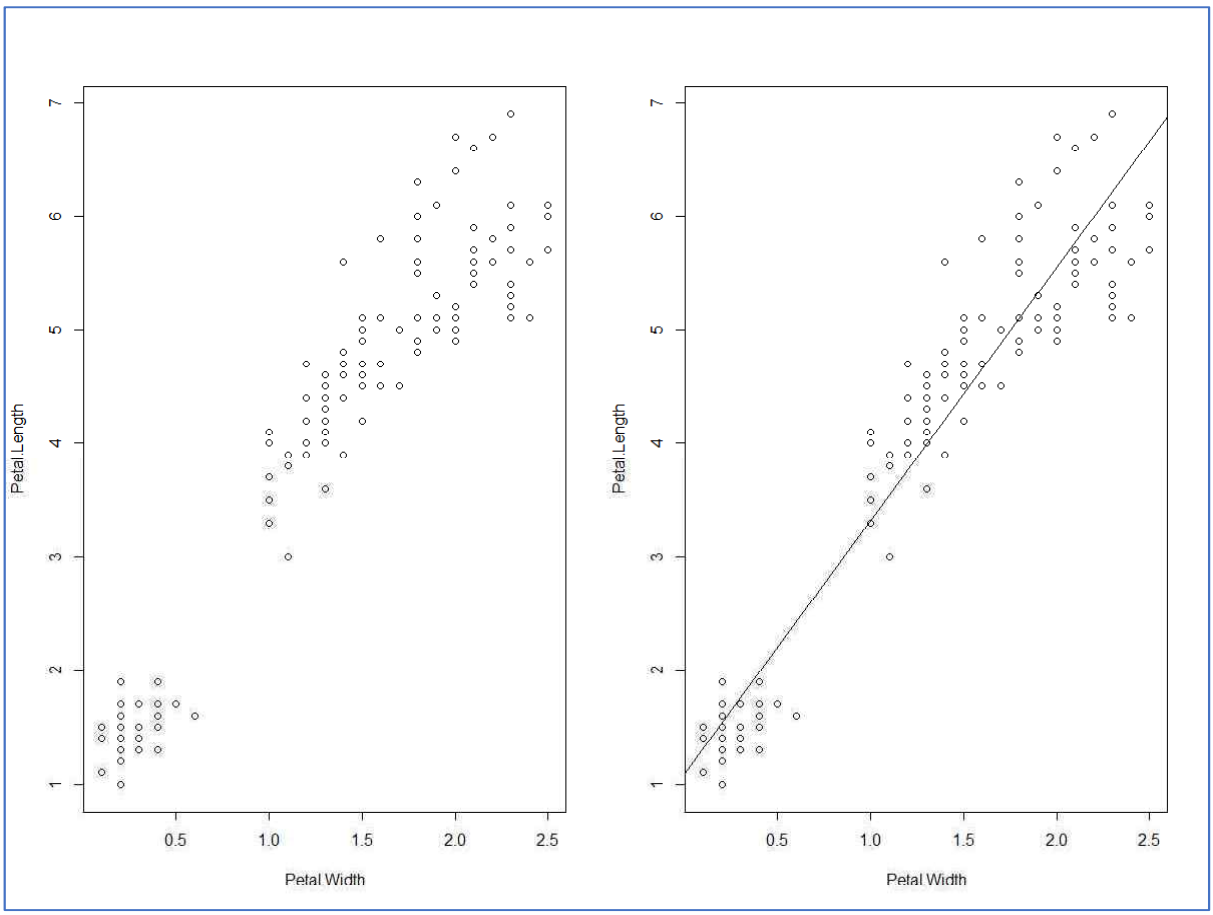
#이는 즉, Petal.Width 가 Petal.Length 를 설명하는데 유의함을 말한다.

<문제 해설>

1) attach 함수를 통해서 iris 내의 데이터에 보다 더 쉽게 접근하도록 한다.

2) Plot(Petal.Width, Petal. Length)를 통해서 x 축에 Petal.Width 의 value 를 Y 축에 Petal. Length 를 갖는 plot 을 그린다. 1)번에서 attach 함수를 사용하지 않았다면, plot(iris$Petal.Width, iris$Petal.Length)라고 입력해야 한다.

3) 아래와 같은 plot 을 확인할 수 있는데, 이를 통해서 Petal.Width 와 Petal.Length 는 서로 선형관계가 있음을 알 수 있다.



<좌측 그림: plot 함수를 통해 생성한 plot, 우측 그림: abline 을 추가한 plot>

4) 회귀분석을 시행하기 위해서 lm 함수를 이용한다. lm(Petal.Length~Petal.Width)에서 반응변수를 먼저 입력하고 그 후에 설명변수를 입력한다. 이 결과를 result 오브젝트에 저장한 후 다시 result 를 입력함으로 아래와 같은 결과를 볼 수 있다. 이를 통해서 적합된 회귀식의 절편은 1,084 이고, 기울기는 2.230 임을 알 수 있다. 따라서 적합된 회귀식은 다음과 같다.

$$Y_i = 1.084 + 2.230 * x_i$$

```
> result

Call:
lm(formula = Petal.Length ~ Petal.Width)

Coefficients:
(Intercept)  Petal.Width
      1.084        2.230
```

< lm(Petal.Length~Petal.Width)의 결과>

5) 설명변수 Petal.width 가 반응변수 Petal.Length 를 설명함에 있어서 유의한지 검정하기 위해서는 적합된 회귀식의 기울기가 통계적으로 유의미한지 검정해야합니다. lm function 에서 이 검정에 대해서 p-value 를 제공해줍니다. 이를 위해서 summary(result)를 입력해줍니다. 앞선 4)번에서 lm function 의 결과를 result 오브젝트에 저장했으므로 summary(result)라고만 입력해주어도 됩니다. 그 결과는 아래와 같습니다. 즉, Petal.width 의 기울기에 대해서 검정했을 때, p-value 값이 무척 작음을 알 수 있습니다. 이를 통해서 Petal.width 의 기울기가 통계적으로 유의미함을 알 수 있으며, 따라서 Petal.Length 를 설명함에 있어서 Petal.Width 가 유의미함을 의미합니다. P-value 는 2e-16 으로 이는 지수표기법입니다. Ex) 9e+5=900,000

```
> summary(result)

Call:
lm(formula = Petal.Length ~ Petal.Width)

Residuals:
     Min       1Q   Median       3Q      Max
-1.33542 -0.30347 -0.02955  0.25776  1.39453

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.08356    0.07297   14.85   <2e-16 ***
Petal.Width  2.22994    0.05140   43.39   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4782 on 148 degrees of freedom
Multiple R-squared:  0.9271,    Adjusted R-squared:  0.9266
F-statistic:  1882 on 1 and 148 DF,  p-value: < 2.2e-16
```

< summary(result)의 결과>