통계학실험 제 8장

상관분석과 회귀분석

과목명	통계학실험 (009)
담당교수명	정상아
제출일	2016.05.27
학과	공과대학 컴퓨터공학부
학번, 이름	2016-17101, 김종범

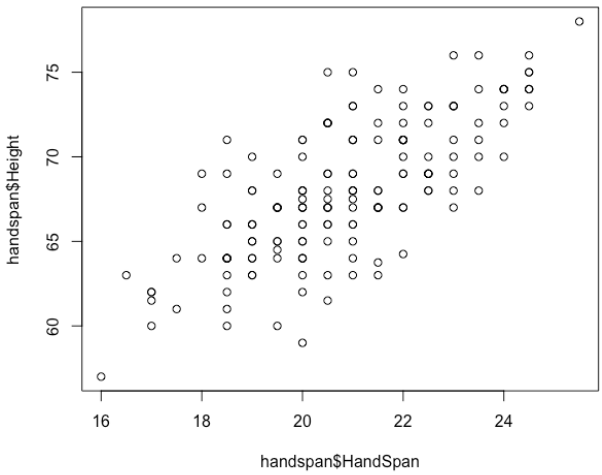
8장 예제 1.

(1)

두 변수의 상관계수는 0.7395375 이다.

산점도는 오른쪽 그림과 같다.

두 변수는 상관관계가 존재하고, 선형적 연관성이 존재한다고 볼 수 있다.



(2)

```
data:  handspan$HandSpan and handspan$Height
t = 14.113, df = 165, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6620252 0.8013971
sample estimates:
      cor
0.7395375
```

상관분석 결과로 p-value가 2.2e-16이므로 유의수준 5%에서 상관관계가 존재한다고 할 수 있다..

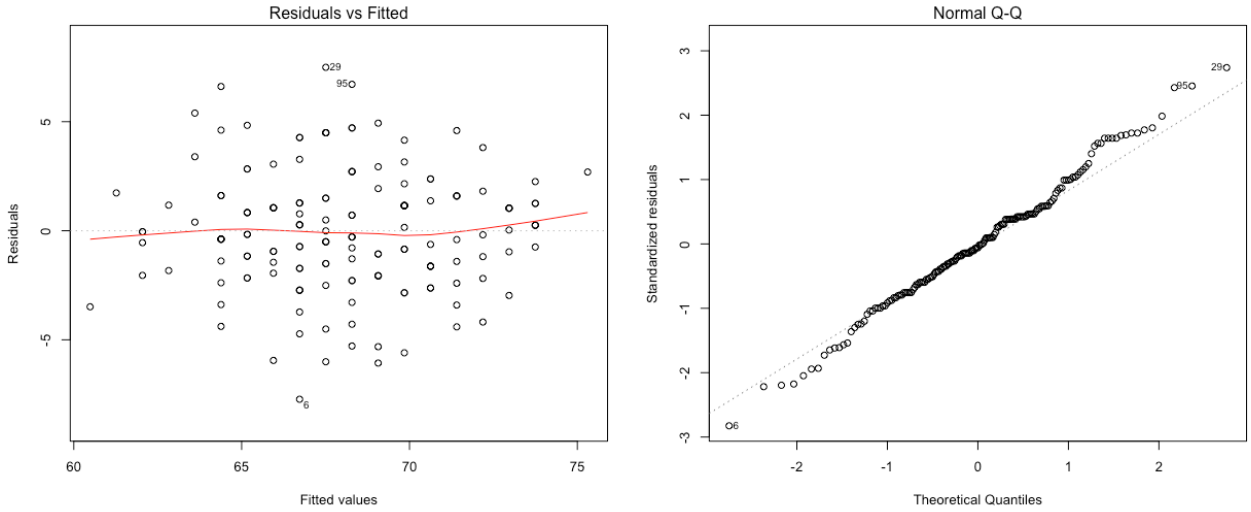
(3)

추정된 회귀식을 y = 1.56x + 35.53 이다.

```
lm(formula = handspan$Height ~ handspan$HandSpan)
Residuals:
    Min      1Q  Median      3Q     Max
-7.7266 -1.7266 -0.1666  1.4933  7.4933
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        35.5250     2.3160   15.34   <2e-16 ***
handspan$HandSpan   1.5601     0.1105   14.11   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 2.744 on 165 degrees of freedom
Multiple R-squared:  0.5469,  Adjusted R-squared:  0.5442
F-statistic: 199.2 on 1 and 165 DF,  p-value: < 2.2e-16
```

회귀 직선 유의성 검사 결과로 p-value가 2.2e-16이므로 유의수준 5%에서 회귀 직선이 유의하다고 볼 수 있다.

(4)



잔차도를 보면 전체적으로 등분산을 이루고, 절댓값이 5이하인 값들로 이루어졌다는 것을 알 수 있고, 잔차의 정규 분위수를 보면 정규분포에 가깝다는 것을 확인 할 수 있다. 고로 단순 선형 회귀모형의 적용은 타당하다.

8장 예제 2.

(1)

두 변수의 상관계수는 0.9355037이고,

```
Pearson's product-moment correlation
data:  car$Speed and car$StopDist
t = 20.68, df = 61, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8952425 0.9606129
sample estimates:      cor 0.9355037
```

상관분석 결과로 p-value가 2.2e-16가 나오므로 상관관계가 존재한 강력한 증거가 된다.

(2)

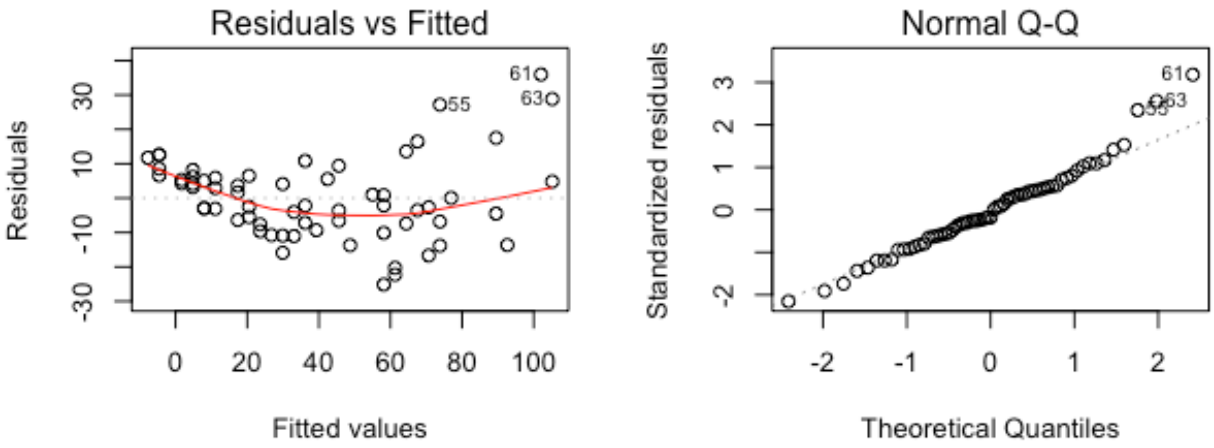
```
lm(formula = car$StopDist ~ car$Speed)
Residuals:
    Min      1Q  Median      3Q     Max
-25.141  -7.300  -2.141   6.044  35.946
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -20.2734     3.2384   -6.26 4.25e-08 ***
car$Speed     3.1366     0.1517   20.68  < 2e-16 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.8 on 61 degrees of freedom
Multiple R-squared:  0.8752,  Adjusted R-squared:  0.8731
F-statistic: 427.7 on 1 and 61 DF,  p-value: < 2.2e-16
```

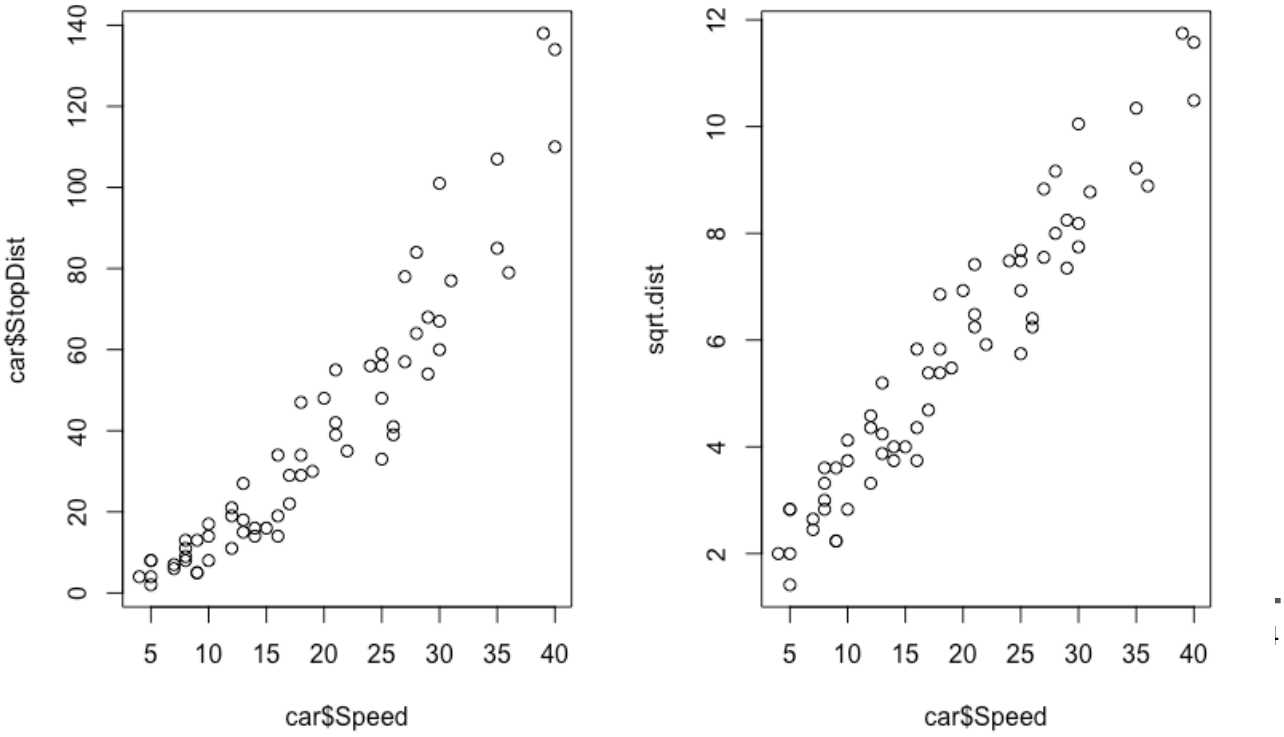
검정 결과로 p-value가 2.2e-16이므로 유의수준 5%에서 모형이 유의하다고 볼 수 있다.

(3)



잔차도를 보면 Residuals의 값이 전체적으로 너무 크다 (5보다 큰 것이 매우 많다). 따라서 단순선형회귀모형의 적용이 타당하다고 볼 수 없다.

(4)



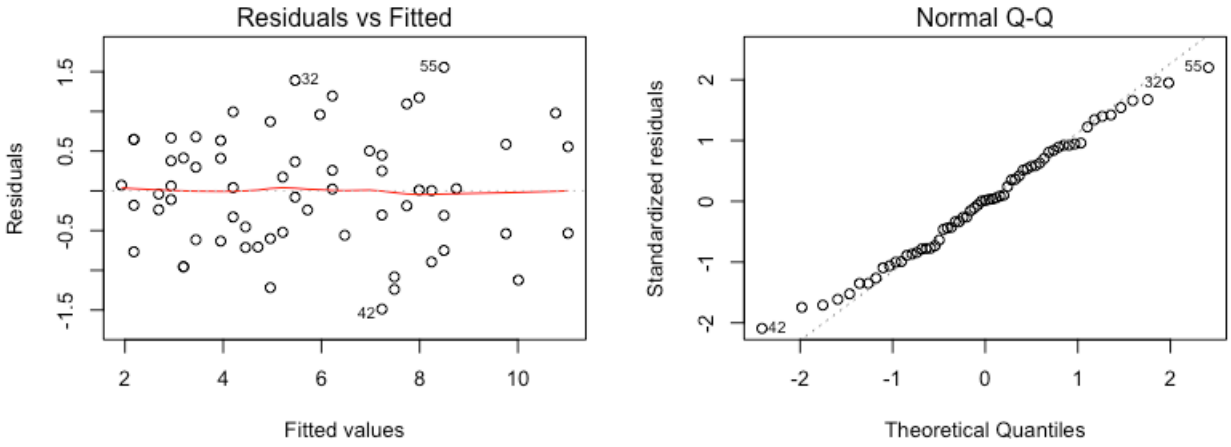
각각의 경우의 산점도는 위 그림과 같다. 새로운 산점도는 기존의 산점도보다 더욱 강력한 직선관계로 볼 수 있다.

(5)

```
lm(formula = sqrt.dist ~ car$Speed)
Residuals:
    Min      1Q  Median      3Q     Max
-1.4879 -0.5487  0.0098  0.5291  1.5545
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.918283   0.197406   4.652 1.82e-05 ***
car$Speed   0.252568   0.009246  27.317  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.7193 on 61 degrees of freedom
Multiple R-squared:  0.9244,  Adjusted R-squared:  0.9232
F-statistic: 746.2 on 1 and 61 DF,  p-value: < 2.2e-16
```

기존의 R-squared값은 0.8752지만, 새로운 모형은 0.9244로 더욱 강력한 회귀선이라는 것을 알 수 있다.

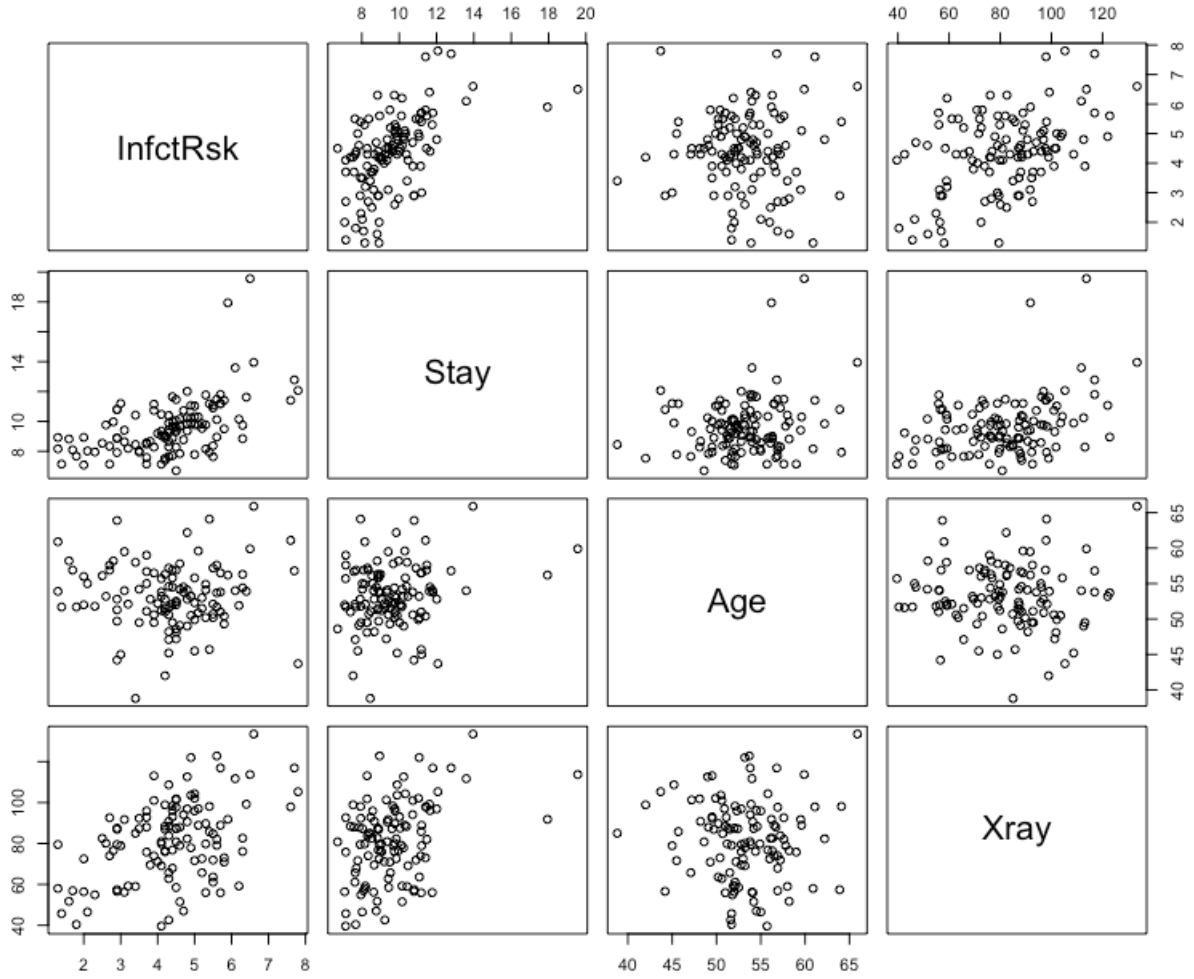
(6)



전체적으로 Residuals의 값이 2이내로 있다는 것을 확인할 수 있었고, 정규분포에 가깝다는 것을 확인할 수 있었다. 즉, 단순선형회귀모형의 적용이 타당하다고 볼 수 있다.

8장 예제 3.

(1)



각각의 산점도는 위의 그림을 통하여 확인 할 수 있다.

```
> cor(newData)
            InfctRsk      Stay          Age       Xray
InfctRsk 1.000000000 0.5334438  0.001093166  0.4533916
Stay     0.533443831 1.0000000  0.188913972  0.3824819
Age      0.001093166 0.1889140  1.000000000 -0.0188549
Xray     0.453391557 0.3824819 -0.018854897  1.0000000
```

각각의 상관계수는 위의 표를 통하여 확인 할 수 있다.

```
> cor.test(newData$InfctRsk, newData$Stay)

     Pearson's product-moment correlation

data:  newData$InfctRsk and newData$Stay
```

```
t = 6.6445, df = 111, p-value = 1.177e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3868338 0.6537511
sample estimates:
      cor
0.5334438

> cor.test(newData$InfctRsk, newData$Age)

     Pearson's product-moment correlation

data:  newData$InfctRsk and newData$Age
t = 0.011517, df = 111, p-value = 0.9908
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1836737  0.1857855
sample estimates:
        cor
0.001093166

> cor.test(newData$InfctRsk, newData$Xray)

     Pearson's product-moment correlation

data:  newData$InfctRsk and newData$Xray
t = 5.3593, df = 111, p-value = 4.585e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2932204 0.5888060
sample estimates:
      cor
0.4533916
```

위의 결과는 각각의 변수들과 상관분석을 한 결과이고, Age는 p-value가 매우 1에 가까우므로 InfectRsk과 상관관계가 없다고 할 수 있다. 나머지 변수들에 관해서는 p-value가 매우 작으므로 상관관계가 있다고 할 수 있다.

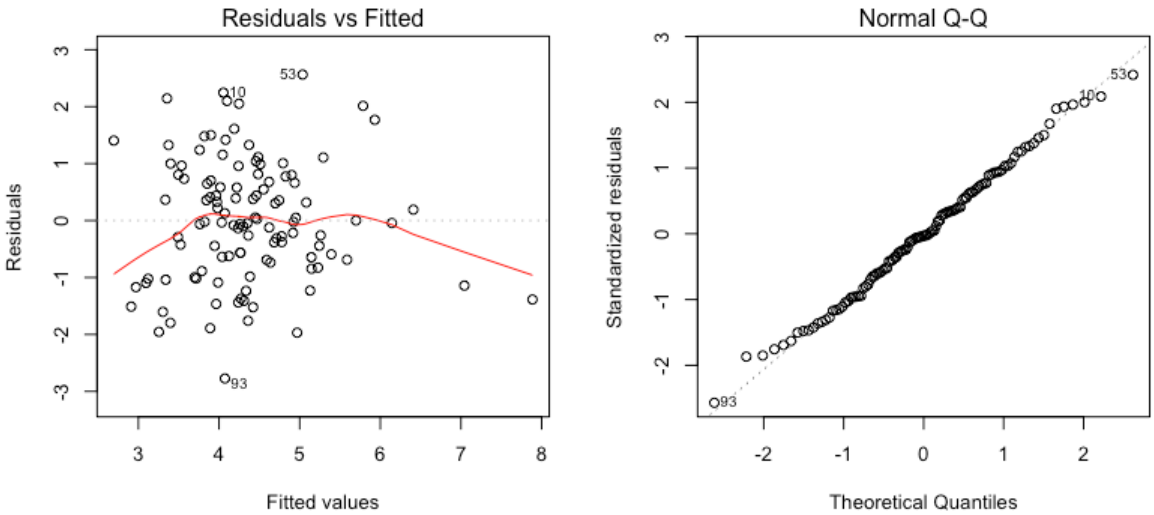
(2)

```
lm(formula = InfctRsk ~ Stay + Age + Xray, data = newData)
Residuals:
     Min       1Q   Median       3Q      Max
-2.77320 -0.73779 -0.03345  0.73308  2.56331
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.001162   1.314724   0.761 0.448003
Stay         0.308181   0.059396   5.189 9.88e-07 ***
Age         -0.023005   0.023516  -0.978 0.330098
Xray         0.019661   0.005759   3.414 0.000899 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 1.085 on 109 degrees of freedom
Multiple R-squared:  0.363,   Adjusted R-squared:  0.3455
F-statistic:  20.7 on 3 and 109 DF,  p-value: 1.087e-10
```

다중선형회귀모형 검정 결과 p-value가 1.087e-10로 나와, 유의수준 5%에서 모형은 유의하다고 볼 수 있다.

결과에서 Age의 p값은 0.33으로 매우 크게 나왔고, 나머지 변수들은 0.001이하의 매우 작은 값을 나타낸다. 즉, Age의 변수는 유의하다고 볼 수 없고, 나머지 변수들은 유의하다고 볼 수 있다.

(3)



전체적으로 Residuals의 값의 절댓값이 3이내인 것을 확인할 수 있고 정규분포에 가깝다는 것을 알 수 있다. 즉, 다중선형회귀모형의 적용은 타당하다고 볼 수 있다.