R컴퓨팅

15강 회귀분석

정보통계학과 장영재 교수

1. 선형관계
2. 단순 선형 회귀모형
3. 중회귀모형
4. 회귀진단
5. 변수선택

1 선형관계

1 선형관계

- 회귀분석에 사용되는 자료는 반응변수(일반적으로 Y로 표기)와 설명변수(일반적으로 X로 표기)가 짝으로 관측되는 경우

짝으로 관측되는 자료가 선형관계를 가지고 있는 경우에
선형 회귀모형을 사용하여 분석

1 선형관계

※ 나이와 키 자료

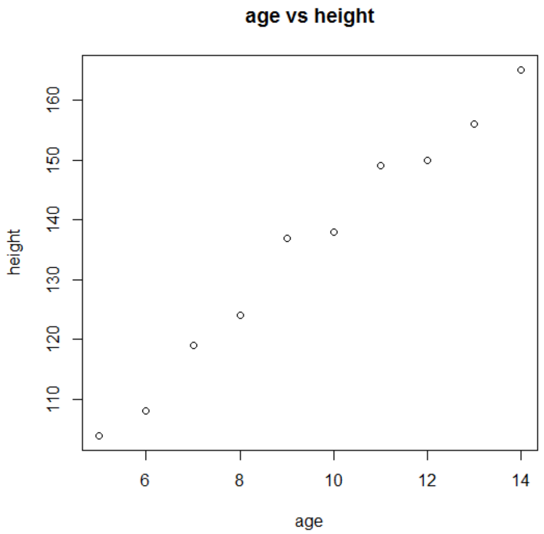
- 자료가 주어진 경우에 처음으로 할 일은 두 변수 사이의 산점도를 그리는 것

나이 (years)	키 (cm)
5	104
6	108
7	119
8	124
9	137
10	138
11	149
12	150
13	156
14	165

1 선형관계

보기 15-1: 위의 자료를 입력하고 산점도 그리기

```
>age<-5:14
>height<-c(104,108,119,124,137,138,149,150,156,165)
>plot(age,height,main="age vs height")
```



2 단순 선형 회귀모형

2 단순 선형 회귀모형

- 두 변수사이에 관계가 선형임을 확인했으므로 선형 회귀모형을 이용하여 분석할 수 있음

- 선형 회귀모형은 반응변수와 설명변수간에 선형관계가 있다고 가정하고 하나의 직선으로 두 변수사이의 관계를 설명하는 모형

- 반응변수는 y, 설명변수는 x라고 할 때,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1 \cdots n.$$

위의 식에서 β_0는 회귀직선의 Y 절편이고, β_1은 기울기임
ε_i는 오차항 (평균이 0이고 분산은 σ^2라고 가정)

2 단순 선형 회귀모형

- 직선들을 비교할 수 있는 척도가 필요한데 가장 많이 사용되는 척도가 잔차제곱합
- 잔차란 관측치에서 예측치를 뺀 값을 의미하고 이 차이들의 제곱합이 잔차제곱합(residual sum of squares, RSS)
- 만약 반응변수 (이 경우에 키)를 y_i , 설명변수를 (이 경우에 나이) x_i라고 하고 회귀모형이 주는 예측치를 $\hat{y}_i$이라고 하면

  잔차 $r_i = y_i - \hat{y}_i$로 정의되고

  잔차 제곱합은 $\sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$으로 나타낼 수 있음

-우리가 원하는 직선은 이 잔차 제곱합을 최소로 하는 선
-잔차제곱합을 최소로 하는 직선을 추정하는 방법이 최소제곱추정법

2 단순 선형 회귀모형

R에서 선형회귀직선을 추정하는 키워드는 “lm” 이고 이는

linear model를 뜻함

```
>lm1<-lm(height~age)
>lm(Y ~ X) #단순선형회귀 모형으로 반응변수는 Y, 설명변수는 X에 저장.
>lm(final ~ midterm, data=grades) # 데이터는 grades라는 데이터 프레임
```

lm 함수를 사용해서 모형을 적합하고 나서 그 결과를 lm1 이라는 object에 저장할 경우 lm1은 다음과 같이 여러 가지 attributes를 가지고 있는 리스트 object가 됨

2 단순 선형 회귀모형

```
> names(lm1)
[1] "coefficients" "residuals" "effects" "rank"
[5] "fitted.values" "assign" "qr" "df.residual"
[9] "xlevels" "call" "terms" "model"
```

보기 15-2: summary() 함수를 이용하여 lm1에 저장된 내용을 출력하고 회귀직선을 도출하기

```
>summary(lm1)
Call:
lm(formula = height ~ age)
Residuals:
Min 1Q Median 3Q Max
-3.3273 -1.6455 -0.5000 0.5864 5.3818
```

2 단순 선형 회귀모형

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  70.7455     3.1831   22.23 1.78e-08 ***
age           6.7636     0.3207   21.09 2.69e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.913 on 8 degrees of freedom
Multiple R-squared:  0.9823,    Adjusted R-squared:  0.9801
F-statistic: 444.7 on 1 and 8 DF,  p-value: 2.685e-08
```

› 회귀직선은 다음과 같음

$$Height = 70.7455 + 6.7636 \times age$$

2 단순 선형 회귀모형

- Std. Error와 t value, 그리고 Pr(>|t|) 값이 주어지는데 특히 마지막 열에 나오는 Pr(>|t|) 값은 각각의 회귀계수가 0인지 아닌지를 검정하는 경우에 계산되는 p-value 값

- Multiple R-squared: 0.9823 라고 나와 있음을 볼 수 있는데 이 값은 Y(반응변수)의 총변동량 중에서 얼마만큼이 모형에 있는 X(설명변수)로 설명되고 있는 지를 나타내주는 값

*R-squared

= 선형모형에 의해서 설명되는 변동량 / Y의 총변동량

= 1 - 잔차제곱합/Y의 총변동량

2 단순 선형 회귀모형

- 보기 15-3: lm1 산출을 위해 사용한 데이터와 lm1의 객체 residuals(잔차) 등을 이용하여 R-squared 값을 구하기

```
> sum((height-mean(height))^2)     #반응변수 height의 총변동량
[1] 3842

> sum((lm1$residuals - mean(lm1$residuals))^2)     #잔차제곱합
[1] 67.89091

> rss<-sum((lm1$residuals - mean(lm1$residuals))^2)
> sst<-sum((height-mean(height))^2)
> 1-rss/sst
[1] 0.9823293
```

2 단순 선형 회귀모형

- 단순선형회귀모형에서 이 R-squared값은 설명변수와 반응변수사이의 상관계수 값을 제곱한 것과 같음

```
> cor(age,height)
    [1] 0.9911253
    > cor(age,height)^2
    [1] 0.9823293
```

2 단순 선형 회귀모형

- 보기 15-4: 벡터 (6.3,7.2,10.5,13.6)를 age2에 저장한 뒤, 이를 age.new라는 이름의 데이터프레임으로 생성하고 predict() 함수와 lm1 회귀직선을 이용하여 새로운 나이 자료 age.new에 관한 예측값 구하기Y

2 단순 선형 회귀모형

```
> age2<-c(6.3,7.2,10.5,13.6)   #예측에 사용될 새로운 자료
> age.new<-data.frame(age2) #dataframe을 새로 생성한다.
> colnames(age.new)<-"age"   #예측에 사용될 변수의 이름을 정해준다.
> age.new
   age
1  6.3
2  7.2
3 10.5
4 13.6
> predict(lm1,age.new)
       1        2        3        4
113.3564 119.4436 141.7636 162.7309
```

3 중회귀모형

3 중회귀모형

- 일반적으로 많은 실제 자료들은 2개 이상의 설명변수를 가지고 있는 경우에도 선형모형을 사용해서 회귀분석을 하는 것이 가능하며 이렇게 설명변수가 2개 이상인 경우를 중회귀모형이라고 함

- R에서 제공하는 mtcars 데이터를 출력하고 산점도행렬을 그려보면 다음과 같음

```
> dim(mtcars)
[1] 32 11
> head(mtcars)
```

3 중회귀모형

mpg cyl disp hp drat wt qsec vs am gear carb
Mazda RX4 21.0 6 160 110 3.90 2.620 16.46 0 1 4 4
Mazda RX4 Wag 21.0 6 160 110 3.90 2.875 17.02 0 1 4 4
Datsun 710 22.8 4 108 93 3.85 2.320 18.61 1 1 4 1
Hornet 4 Drive 21.4 6 258 110 3.08 3.215 19.44 1 0 3 1
Hornet Sportabout 18.7 8 360 175 3.15 3.440 17.02 0 0 3 2
Valiant 18.1 6 225 105 2.76 3.460 20.22 1 0 3 1

3 중회귀모형

※ mtcars
자료의 산점도
→

3 중회귀모형

- 보기 15-5: R에서 제공하는 mtcars 자료를 이용하여 중회귀분석을 실행하기

- R에서 중회귀모형을 사용하여 모형을 적합하기

```
> lm.cars<-lm(mpg~., data=mtcars)
> summary(lm.cars)
Call:
lm(formula = mpg ~ ., data = mtcars)
Residuals:
Min 1Q Median 3Q Max
-3.4506 -1.6044 -0.1196 1.2193 4.6271
```

3 중회귀모형

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.30337 18.71788 0.657 0.5181
cyl –0.11144 1.04502 –0.107 0.9161
disp 0.01334 0.01786 0.747 0.4635
hp –0.02148 0.02177 –0.987 0.3350
drat 0.78711 1.63537 0.481 0.6353
wt –3.71530 1.89441 –1.961 0.0633 .
qsec 0.82104 0.73084 1.123 0.2739
vs 0.31776 2.10451 0.151 0.8814
am 2.52023 2.05665 1.225 0.2340
gear 0.65541 1.49326 0.439 0.6652
carb –0.19942 0.82875 –0.241 0.8122
---
```

3 중회귀모형

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

Residual standard error: 2.65 on 21 degrees of freedom

Multiple R-squared: 0.869, Adjusted R-squared: 0.8066

F-statistic: 13.93 on 10 and 21 DF, p-value: 3.793e-07

4 회귀진단

4 회귀진단

- 회귀진단은 기본적으로 주어진 자료가 선형모형을 사용하기에 얼마나 적합한 지를 알아보는 것

- 선형모형의 기본적인 가정

① 반응변수와 설명변수는 선형관계를 가지고 있음
② 오차항은 평균이 0이고 등분산을 가짐

- 진단을 위한 좋은 방법은 잔차와 설명변수 사이의 산점도를 보는 것

4 회귀진단

- 보기 15-6: 다음의 자료를 가지고 산점도를 만들고 상관계수를 산출한 뒤 선형회귀모형을 적합하기

```
>x<-c( 1, 4, 17, 30, 40, 49, 54, 60, 63, 78)
>y<-c(3, -52, -1116, -3535, -6316, -9500, -11551, -14274, -
15745, -24173)
> cor(x,y)
[1] -0.9584267
```

- 두 변수 사이의 상관계수는 -0.958로 상당히 강한 음의 상관관계를 가지고 있어서 선형모형을 이용해도 괜찮아 보임

```
> lm1<-lm(y~x)
> summary(lm1)
```

4 회귀진단

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2973.48    1438.75   2.067   0.0726 .
x            -292.91      30.83  -9.500 1.24e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2407 on 8 degrees of freedom
Multiple R-squared:  0.9186,    Adjusted R-squared:
0.9084
F-statistic: 90.26 on 1 and 8 DF,  p-value: 1.243e-05
```

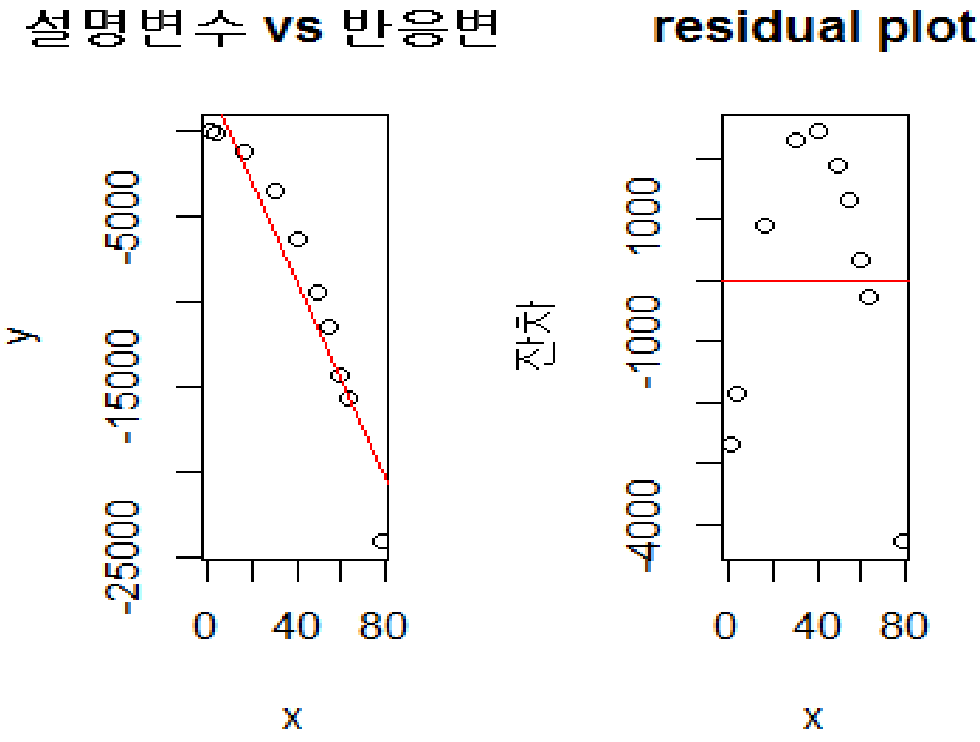
4 회귀진단

설명변수와 반응변수 사이의 산점도와 설명변수와 잔차 사이의 산점도

```
>par(mfrow=c(1,2))     # 그림판을 2개의 열로 나눔
>plot(x,y,main="설명변수 vs 반응변수")
>abline(lm1,col=2)     #회귀직선을 빨간색으로 그리기
>plot(x,lm1$residual,ylab="잔차", main="residual plot")
>abline(h=0,col=2)     #y=0 선을 그려서 비교를 용이하게 함
```

4 회귀진단

※ 자료의 산점도와 잔차 산점도



4 회귀진단

- 왼쪽의 산점도를 보면 문제가 있어 보이지는 않지만 오른쪽에 있는 잔차 산점도를 보면 잔차들이 크게 휘어진 곡선이므로 설명변수와 반응변수의 관계가 선형이 아님을 의미

*실제로 이 데이터는 반응변수가 설명변수의 2차 제곱항을 가지고 있는 비선형 관계

5 변수선택

5 변수선택

- 설명변수의 수가 많은 경우에 어떻게 최적의 변수를 선택하는지에 대한 방법을 살펴보기로 함

- 좋은 회귀모형이란 다음의 두 가지 관점을 모두 만족시키는 것이 좋음

① 예측 오차가 작아야 함. 즉, 관측치와 예측치의 차이가 작을수록 좋음

② 모형은 간단할수록 좋음

5 변수선택

› 다른 모형들을 비교할 때 가장 많이 사용되는 척도가 AIC(Akaike Information Criterion)

$$AIC = nlog(RSS/n) + 2P.$$

- RSS는 잔차 제곱합(Residual Sum of Squares)이고 n은 관측치의 개수, P는 모형에 포함된 설명변수의 개수

- 사용 가능한 모든 설명변수 중에서 이 AIC 값을 최소화하는 설명변수의 집합이 무엇인가를 찾는 것

› 모든 가능한 경우를 계산하는 것은 현실적으로 어려운 경우가 많으므로 stepwise regression 알고리즘을 이용해서 AIC값이 가장 작은 모형을 찾게 됨

- stepwise regression 방법론은 각각의 스텝에서 설명변수를 하나씩 모형에 포함하거나 하나씩 제거해 가면서 AIC 값이 가장 작은 모형을 찾는 것

5 변수선택

- 보기 15-7: 앞 절에서 사용한 mtcars 자료를 이용해서 stepwise regression 모형을 적합하기(R에서 stepwise regression 방법론은 step() 함수를 이용)

```
> lm1<-lm(mpg~., data=mtcars)
> lm1.step<-step(lm1,direction="both")
Start:  AIC=70.9
mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am +
gear + carb
```

5 변수선택

Df Sum of Sq RSS AIC

– cyl 1 0.0799 147.57 68.915

– vs 1 0.1601 147.66 68.932

– carb 1 0.4067 147.90 68.986

– gear 1 1.3531 148.85 69.190

– drat 1 1.6270 149.12 69.249

– disp 1 3.9167 151.41 69.736

– hp 1 6.8399 154.33 70.348

– qsec 1 8.8641 156.36 70.765

〈none〉 147.49 70.898

– am 1 10.5467 158.04 71.108

– wt 1 27.0144 174.51 74.280

5 변수선택

- 위의 R 결과에서 보면, 처음에는 모든 설명변수를 가지고 있는 full model에서 시작하며 이때의 AIC 값은 70.9 임을 알 수 있음

- 그 다음 스텝에서는 각각의 설명변수를 하나씩 모형에서 빼면서 AIC 값을 계산하는데 예를 들어 cyl 변수를 모형에서 제거한 경우에 AIC 값이 68.915로 최소가 된다면 이 변수를 모형에서 제거

- 이러한 과정을 반복하여 최종 선택된 모형은 다음과 같음

5 변수선택

```
> summary(lm1.step)
Call:
lm(formula = mpg ~ wt + qsec + am, data = mtcars)
Residuals:
Min 1Q Median 3Q Max
-3.4811 -1.5555 -0.7257 1.4110 4.6610
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.6178 6.9596 1.382 0.177915
wt -3.9165 0.7112 -5.507 6.95e-06 ***
qsec 1.2259 0.2887 4.247 0.000216 ***
am 2.9358 1.4109 2.081 0.046716 *
---
```

5 변수선택

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 2.459 on 28 degrees of freedom
Multiple R-squared: 0.8497, Adjusted R-squared: 0.8336
F-statistic: 52.75 on 3 and 28 DF, p-value: 1.21e-11

- 최종적으로 선택된 변수는 wt(무게), qsec(1/4mile 도달하는데 걸린 시간), 그리고 am(automatic/manual transmission)임

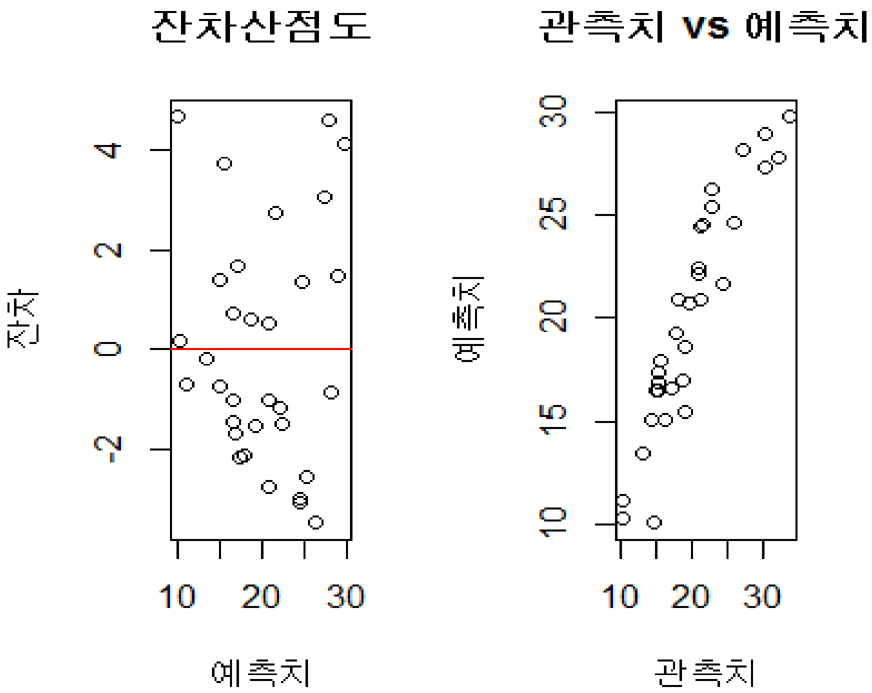
5 변수선택

최종모형의 잔차산점도와 관측치 vs 예측치 산점도

```
> par(mfrow=c(1,2))
>plot(lm1.step$fitted.values, lm1.step$resid,main="잔차산점도",
+ xlab="예측치", ylab="잔차")
> abline(h=0,col=2)
> plot(mtcars$mpg, lm1.step$fitted.values, main="관측치 vs 예측치",
+ xlab="관측치", ylab="예측치")
> abline(a=0,b=1,col=2)
```

5 변수선택

※ 최종모형의 잔차 산점도와 관측치 vs 예측치 산점도



5 변수선택

- 위의 그림에서 볼 수 있듯이 잔차산점도는 0근방에서 아무런 곡선의 패턴도 찾기 힘들고
- 오른쪽에 있는 관측치와 예측치의 산점도를 보면 대부분의 자료가 y=x 선상 근방에 위치하고 있음

- 따라서 이 최종모형은 간단하면서도 예측력이 높고 선형 모형의 가정을 대부분 만족시키는 좋은 모형이라고 할 수 있음

R컴퓨팅

한 학기동안 수고하셨습니다