2020학년도 ( 1 )학기 과제물(온라인제출용)

**교과목명 : 회귀모형**

**학 번 : 201835-368393**

**성 명 : 여준영**

**연 락 처 : 010-5196-1122**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

o 과제유형 : ( ) 형

o 과 제 명 : 1학기 회귀모형 출석수업대체시험 실험실습 과제

- 이하 과제 작성

※ A4용지 편집 사용

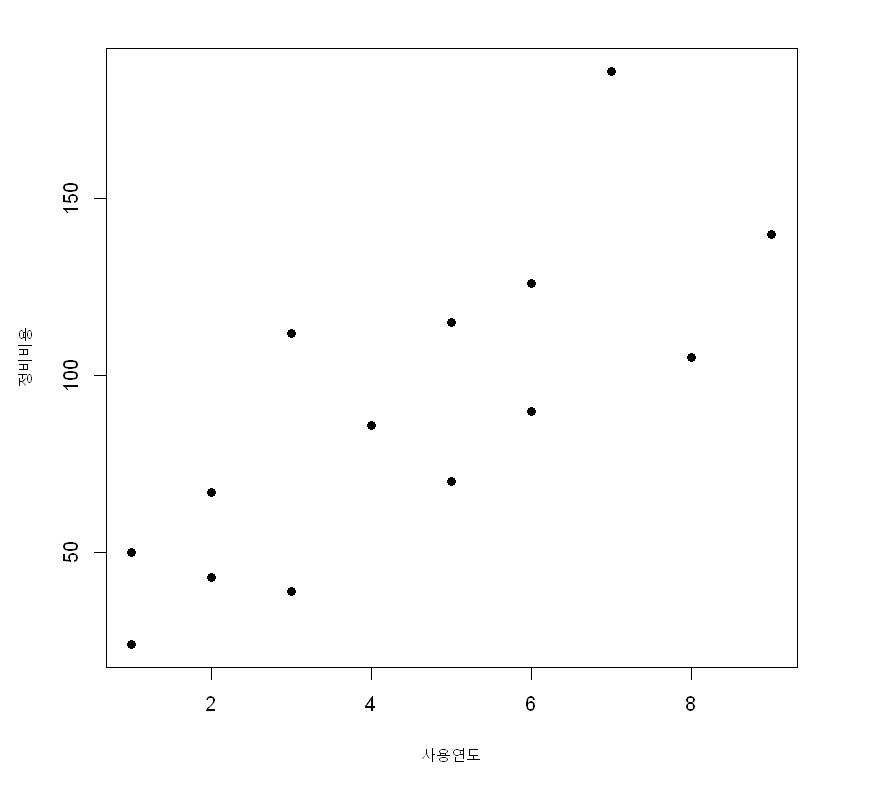
**1장**

1. (1) 이 데이터의 산점도를 그려라
2. 데이터 입력

cost = read.table('maintenance\_cost.csv', header=T, sep = ',')

1. 산점도 그리기

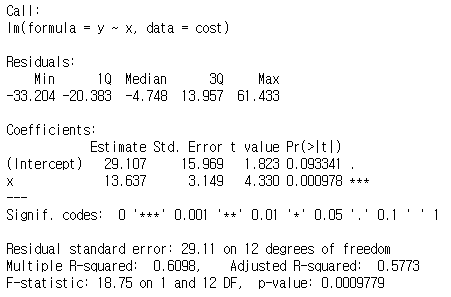
plot(cost$x, cost$y, xlab='사용연도', ylab='정비비용', pch=19)



(2) 최소제곱법에 의한 회귀직선을 적합시켜라.

cost.lm <- lm(y ~ x, data=cost)

summary(cost.lm)



위의 회귀적합 결과에서 회귀계수의 추정갑은 절편 b0 = 29.107이고, 기울기 b1= 13.631으로서 단순 회귀 방정식은

**Y = 29.107 + 13.637 \* X**

가 된다.

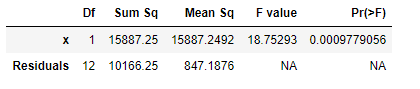
(3) 추정치의 표준오차 Sy.x 를 구하라.

위의 회귀적합결과에서 Residual standard error를 말하며 **29.11**이 된다.

(4) 결정계수와 상관계수를 구하라.

결정계수 : 역시 위의 회귀적합결과에서 Multiple R-squared 값을 보면 0.6098임을 알 수 있다.  
상관계수 : 결정계수는 상관계수의 제곱근 값으로 표현되는데, 적합된 단순회귀직선 방정식이 양의 기울기를 갖고 있으므로 양의 제곱근 값이다. 따라서, **0.7809**가 된다.

(5) 분산분석표를 작성하고 회귀직선의 유의 여부를 검정하라.   
anova(cost.lm)



가 도출된다. 따라서 분산분석표는 아래와 같이 만들 수 있다.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 요인 | 자유도 | 제곱합 | 평균제곱 | F0 | Pr(F>F0) |
| 회귀 | 1 | 15887.25 | 15887.2492 | 18.75293 | 0.0009779056 |
| 잔차 | 12 | 10166.25 | 847.1876 |  |  |
| 계 | 13 | 26053.5 |  |  |  |

여기서 유의수준 a=0.05 수준에서 F-기각역(F, 12;0.05)의 값은  
 qf(0.95, 1, 12)  
> 4.747  
이므로 “F0=18.75293 > F1(1, 12;0.05) =4.747” 이므로 귀무가설 H0:b1=0을 기각한다. 따라서 구해진 회귀직선은 유의하다.

(6) 사용연도가 4년인 기계의 평균정비비용은 어느 정도인가를 추정하라.  
위의 회귀방정식 **Y = 29.107 + 13.637 \* X** 에서 X=4를 대입하면 83.655 로 나오며 단위가 1000원이므로 평균정비비용은 **83,655**원으로 추정된다.

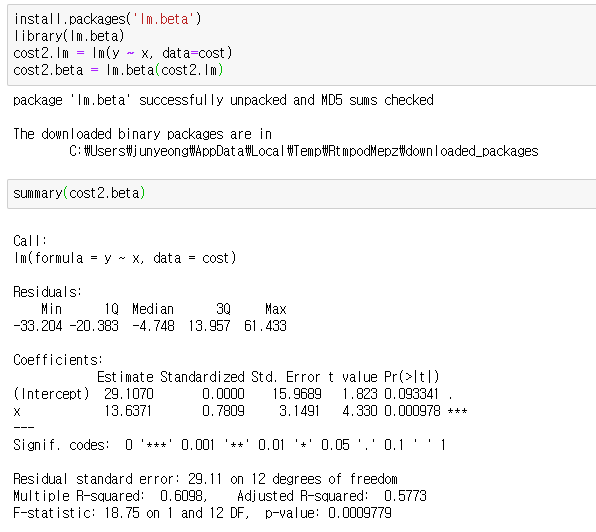
(7) cost.lm$resid  
명령어를 통해 잔차들의 항별 값을 알 수 있으며,   
sum(cost.lm$resid)

명령어를 통해   
-1.06581410364015e-14  
라는 값이 나왔고 잔차의 합이 0임을 알 수 있다.

(8) 잔차의 성질에 의해 0, 그리고 sum(cost$x \* cost.lm$resid) 에 의해서도2.04281036531029e-14, 즉 0이 나온다.

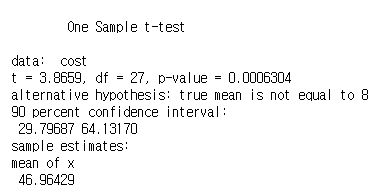
(9) 잔차의 성질에 의해 0, 그리고 sum(cost.lm$fitted \* cost.lm$resid) 에 의해서도 3.41060513164848e-13, 즉 0이 나온다.

(10)



표준화된 변수의 회귀직선을 구하기 위하여 위와 같이 lm.beta 패키지를 이용하였고, 그 결과 y=0.7809 \* z1 이라는 값이 나왔으며 상관계수 r = 0.78 이 나왔다. 따라서 표준화 전의 변수와 상관계수가 같음을 알 수 있다.

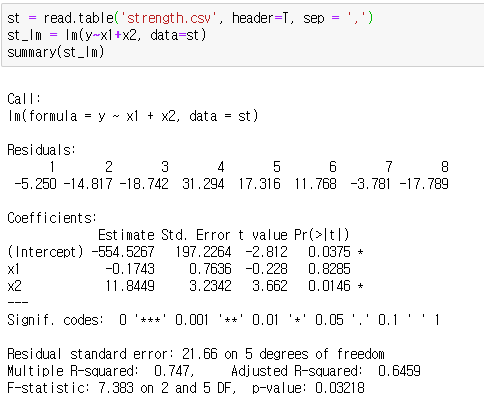
5. t.test(cost, mu = 8, conf.level = 0.9) 로 하면

  
위와 같은 값이 출력되고, p-value가 0.0006304 이므로 0.1보다 작기 때문에 귀무가설을 기각한다.

2장

2.

(1)



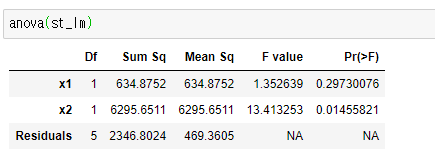
위의 통계량으로 추정된 회귀식은 아래와 같다.  
y = -554.5267 -0.1743x1 + 11.8849x2

(2) MSE는 오차분산의 불편추정량이기 때문에 MSE의 제곱근, 즉 Residual standard error 의 값은 21.66이다. 또한 위의 통계량으로 보아 각 Std.Error 제곱이 분산 추정값이기 때문에 Var(b0) 는 38.898.25, Var(b1)는 0.54, Var(b2) 는 10.46이다.

(3) y = -554.5267 -0.1743x1 + 11.8849x2 에서 각 변수를 대입하면 109.4624 가 나오는 것을 알 수 있다.

(4) 추정된 회귀계수 b1과 b2의 의미는 각각 변수 온도와 압력이 변수 Y에 미치는 정도를 의미한다.

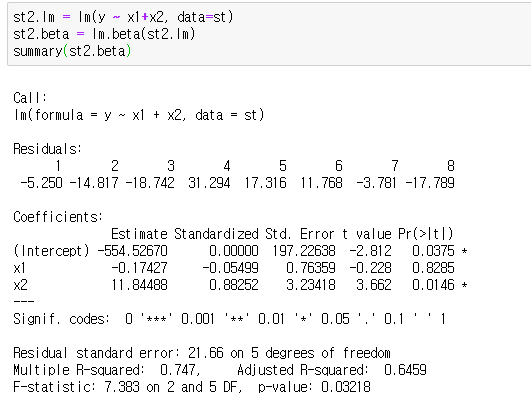
(5)



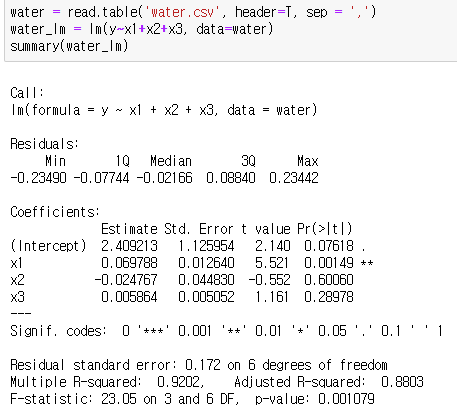
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 요인 | 자유도 | 제곱합 | 평균제곱 | F0 |
| 회귀 | 2 | 6930.6 | 3465.3 | 7.383 |
| 잔차 | 5 | 2346.8 | 469.36 |  |
| 계 | 7 | 9277.4 |  |  |

(6) 결정계수는 위의 통계량에서 Multiple R-squared인 0.747이 된다.

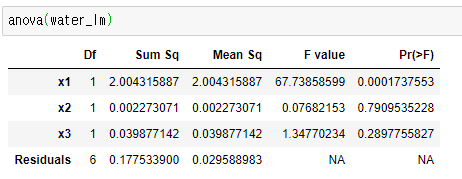
(7)

  
lm.beta 패키지를 사용하여 구해보면   
y= -0.05499 \* z1 + 0.88252\* z2 이라는 표준화된 중회귀직전방정식을 도출할 수 있다.

1. (1)

  
위의 통계량을 바탕으로 y=2.409 + 0.0698x1 – 0.024767x2 + 0.0058x3 라는 회귀직전 방정식을 도출할 수 있다. 그런데 변수 x2의 t-value가 -0.552이고 Pr값이 0.6이므로 a=0.05에서 귀무가설을 기각하지 못하므로 변수 x2는 y를 설명하는데 영향을 준다고 할 수 없다고 할 수 있다.

(2) 추정된 회귀계수 b1, b2, b3의 의미는 각각 변수 x1, x2, x3가 변수 y에 미치는 정도를 의미한다.

(3)   


|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 요인 | 자유도 | 제곱합 | 평균제곱 | F0 |
| 회귀 | 3 | 2.046 | 0.682 | 23.05 |
| 잔차 | 6 | 0.178 | 0.030 |  |
| 계 | 9 | 2.224 |  |  |

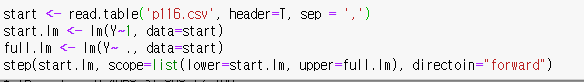
결정계수는 Multiple R-squared 값으로 0.9202 이다.

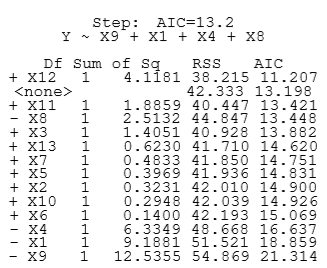
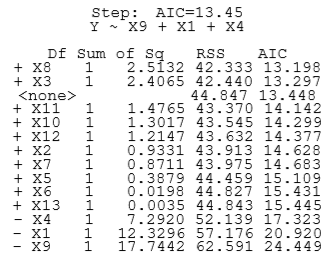
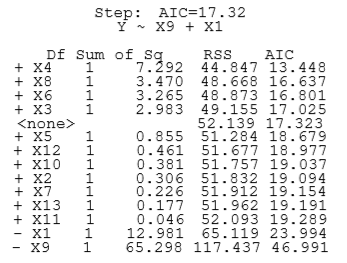
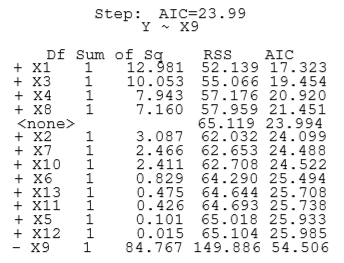
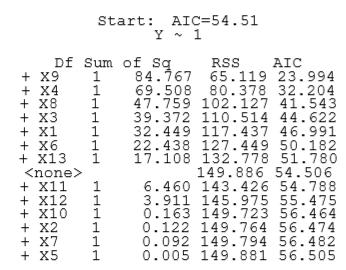
(4) 위의 추정된 회귀직선방정식 y=2.409 + 0.0698x1 – 0.024767x2 + 0.0058x3 각각 x1=20, x2=27, x3=60을 대입해보면, 3.4842라는 y값이 나오고, 평균 물 소비량은 3484.2톤이 된다.

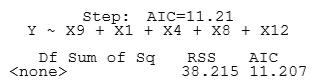
**3장**

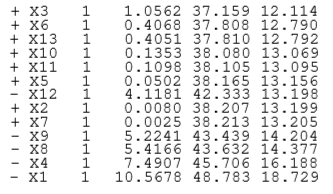
3번

(1)





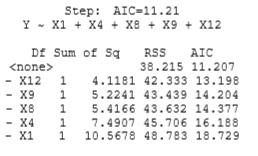
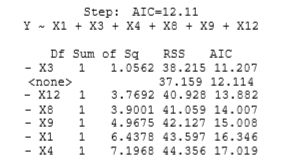
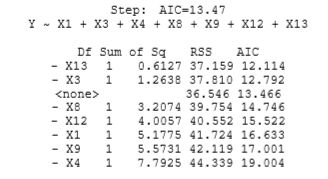
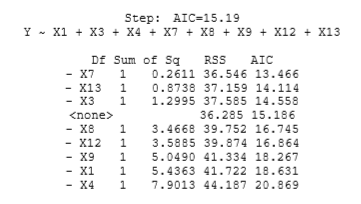
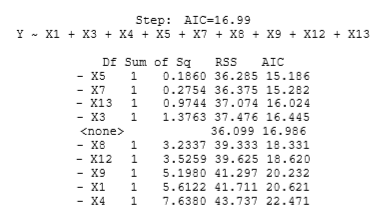
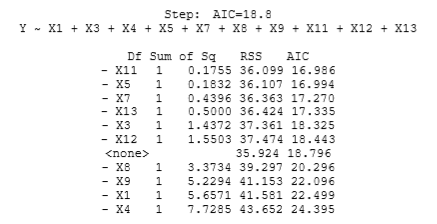
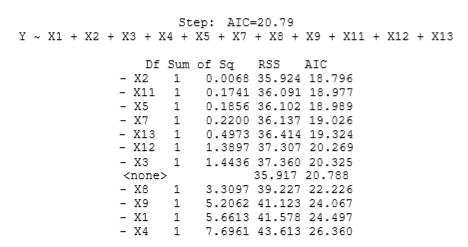
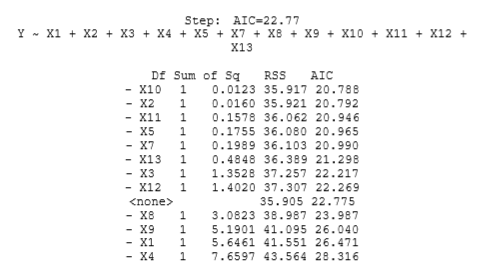
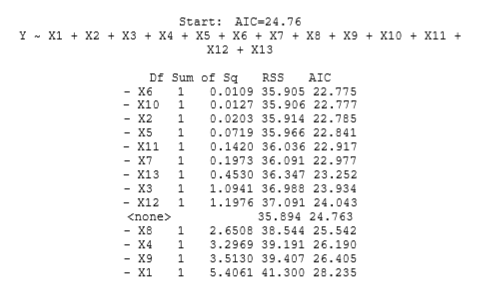




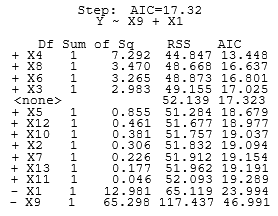
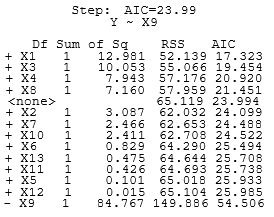
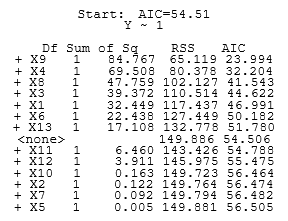
위의 결과를 보면 AIC값이 가장 작은 설명 변수로 X9가 선택되었다. 그 이후의 스텝에서 X1, X4, X8, X12 가 선택되어 최종적으로 중단되었다. 따라서 X9, X1, X4, X8, X12를 포함하는 모형이 최적회귀모형으로 결정된다.

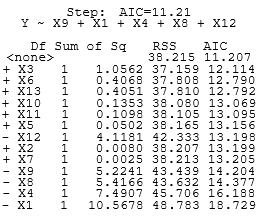
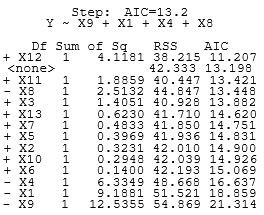
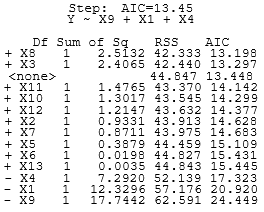
(2)



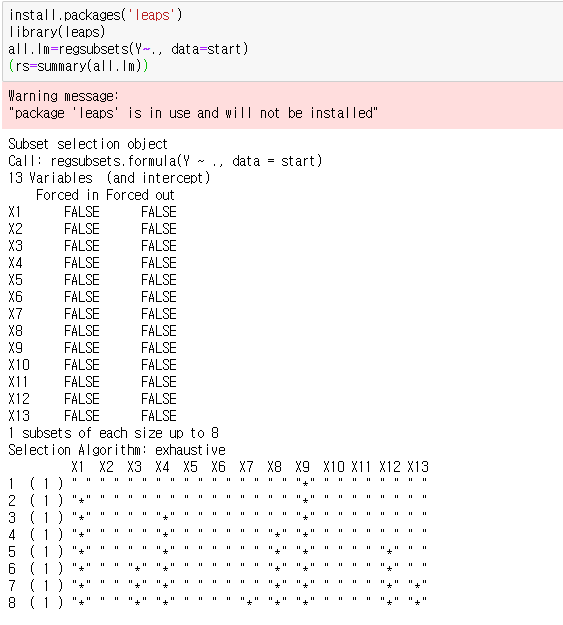
  
에서 고려대상인 13개의 설명변수 모두를 포함하는 다중회귀모형에서 순서대로 X6, X10, X2, X11, X5, X7, X13, X3 이 제거되고 5개의 설명변수(X12, X9, X8, X4, X1)를 포함하는 모형이 최적회귀모형으로 결정되었다.

(3)   

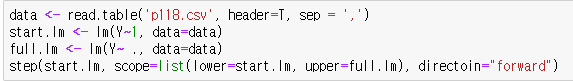
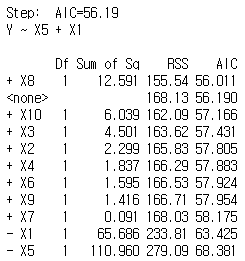
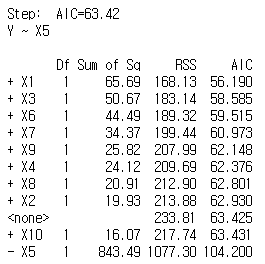
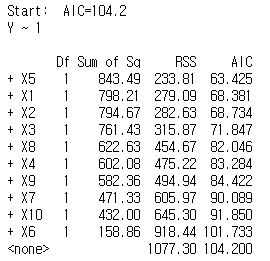


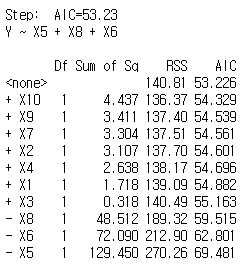
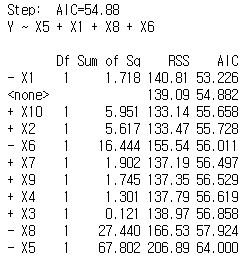
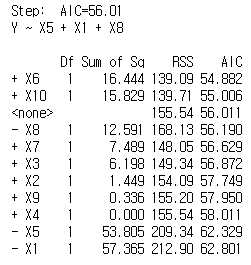
위의 출력결과를 바탕으로 최종 모형으로 선택된 모형은 X9, X1, X4, X8, X12임을 알 수 있다.

(4)   
  
위의 결과를 바탕으로 모든 가능한 회귀에서 선택된 변수는 X1, X3, X4, X7, X8, X9, X12, X13이 된다.

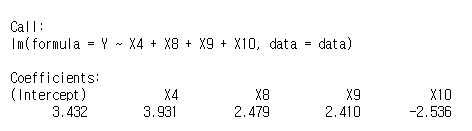
(5) 네가지를 비교해봤을 때, 앞, 뒤, 단계별 회귀방법에서 공통의 결과로서 나온 X9, X1, X4, X8, 12를 선택해봄직하다.

**4번**

(1)   


  
위의 결과를 보았을 때 AIC값이 가장 작은 설명변수로 먼저 X5가 선택되었다. 그 이후에 X1, X8, X6이 선택되었지만 X1이 기각되면서 최종적으로 X5, X8, X6이 선택되었다.

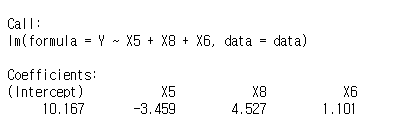
(2)   

X4, X8, X9, X10이 선택되었다.

(3)

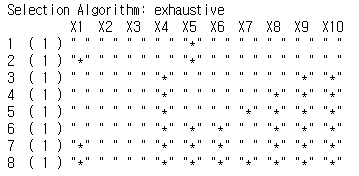




X5, X8, X6이 선택되었다.

(4)



  
최종적으로 X1, X4, X5, X6, X7, X8, X9, X10이 선택되었다.

(5) 문항 3과는 다르게 방법별로 모형의 차이가 약간 존재한다. 앞으로부터 선택법과 단계별 회귀방법은 동일한 결과를 도출했지만, 뒤로부터 제거법과 모든 회귀방법은 모형이 상이하였다.