4.6 $X_1, X_2, \cdots, X_p$에 대한 Y의 회귀에서, 다음 그래프들은 통상적인 최소제곱회귀의 가정들을 평가하기 위해 사용된다.

1. Y 대 각 예측변수 X_j의 산점도
2. $X_1, X_2, \cdots, X_p$의 산점도 행렬
3. 내적 표준화잔차의 정규확률플롯
4. 잔차 대 적합값플롯
5. 잠재성-잔차플롯
6. Cook의 거리에 대한 인덱스플롯
7. Hadi의 영향력 측도에 대한 인덱스 플롯

이들 각 그래프에 대하여

(a) 각 그래프에 의해 확인될 수 있는 가정들은 무엇인가?
(b) 가정들이 위반되지 않는 것으로 보이는 그래프의 예를 그려라.
(c) 가정들이 위반되었음을 나타내는 그래프의 예를 그려라.

Solve)
(a) 각 그래프를 통해, 다음의 가정들을 확인할 수 있다.

1. 모형의 형태에 대한 가정(선형성)
: Y 대 각 예측변수 X_j의 산점도(1)

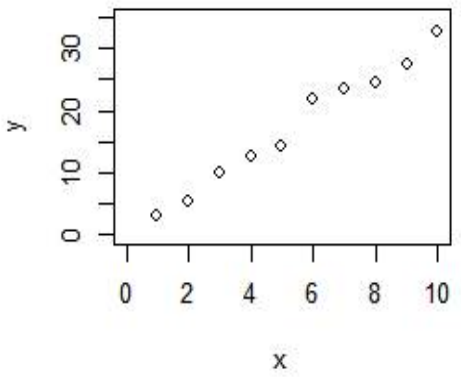
2. 오차에 대한 가정
1) 정규성 : 내적 표준화잔차의 정규확률 플롯(3)
2) 등분산성 : 잔차 대 적합값플롯(4)
3) 독립성 : 없음

3. 예측변수들에 대한 가정(예측변수 사이에는 공선성이 없어야한다.)
: $X_1, X_2, \cdots, X_p$의 산점도 행렬(2)

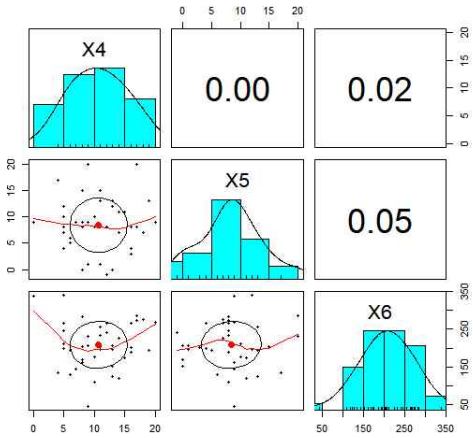
4. 관측개체에 대한 가정(모든 관측개체들은 동일하게 신뢰할 만하며, 결론을 도출함에 있어서 거의 동등한 역할을 한다.)
: 잠재성-잔차플롯(5), Cook의 거리에 대한 인덱스플롯(6)
, Hadi의 영향력 측도에 대한 인덱스 플롯(7)

(b) 다음은 차례대로 (1)~(7)번의 가정이 위반되지 않은 그래프이다.

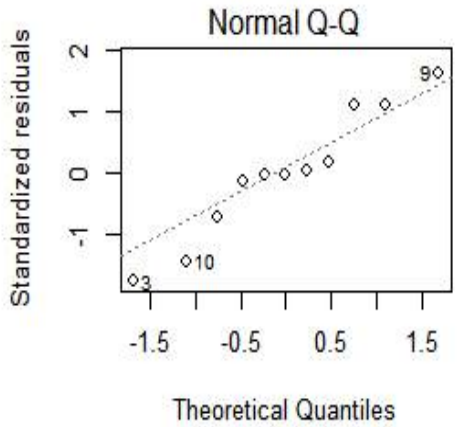
1) Y 대 각 예측변수 X_j의 산점도



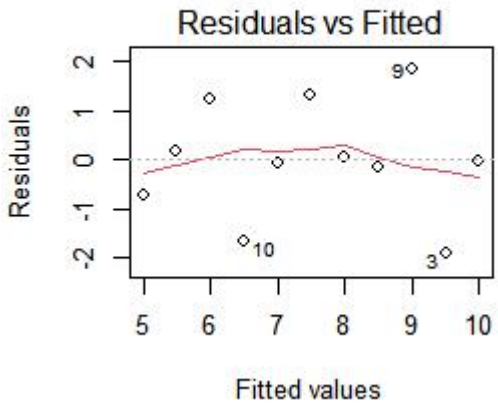
2) $X_1, X_2, \cdots, X_p$의 산점도 행렬



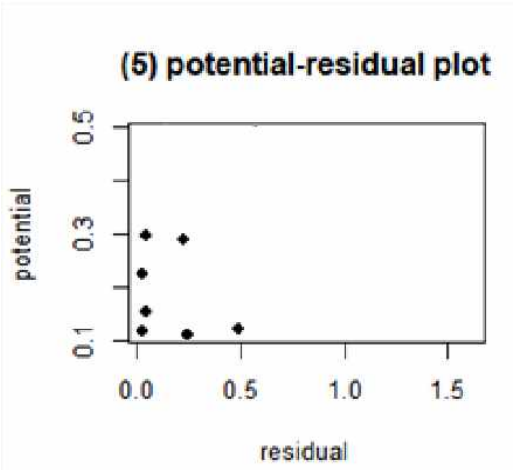
3) 내적 표준화잔차의 정규확률플롯



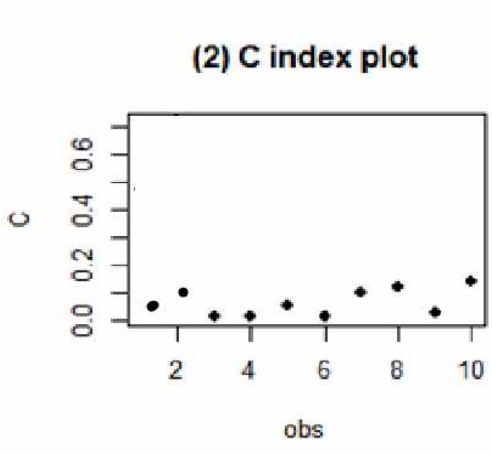
4) 잔차 대 적합값플롯



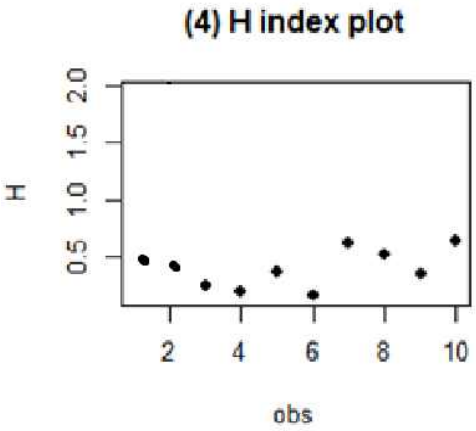
5) 잠재성-잔차플롯



6) Cook의 거리에 대한 인덱스플롯

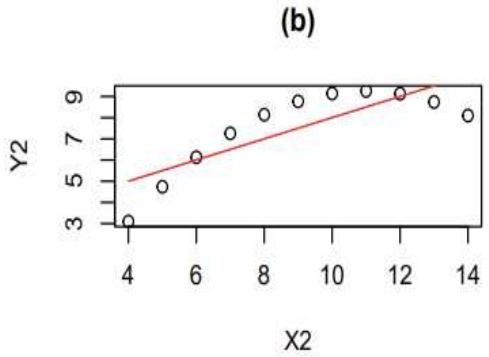


7) Hadi의 영향력 측도에 대한 인덱스 플롯

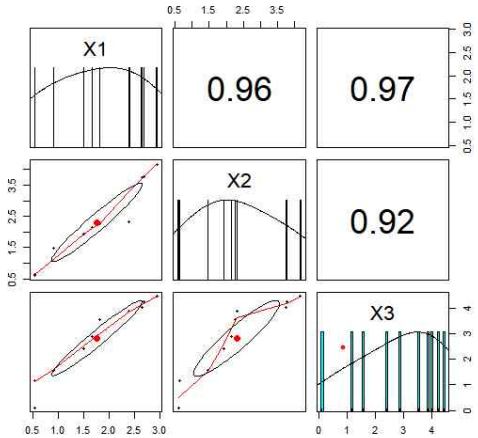


c) 다음은 차례대로 (1)~(7)번의 가정이 위반된 그래프이다.

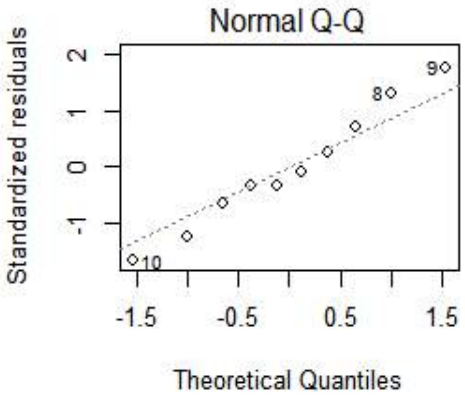
1) Y 대 각 예측변수 X_j의 산점도



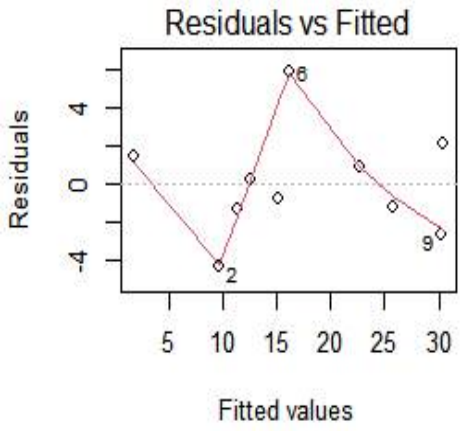
2) $X_1, X_2, \cdots, X_p$의 산점도 행렬



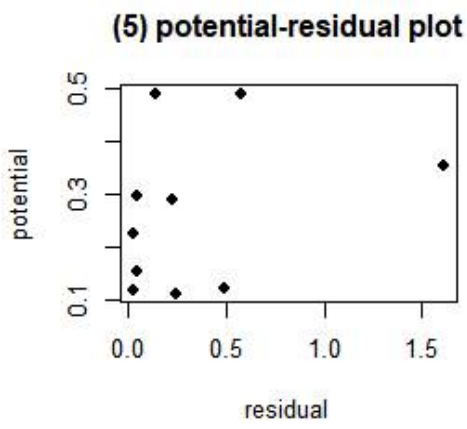
3) 내적 표준화잔차의 정규확률플롯



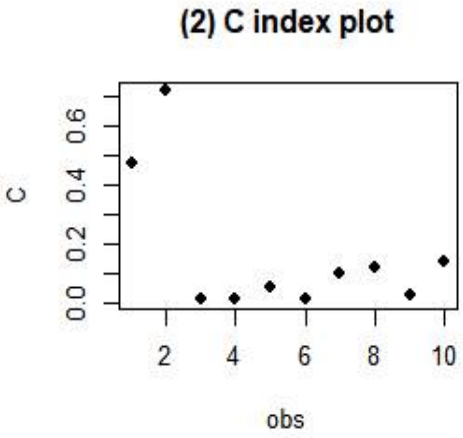
4) 잔차 대 적합값플롯



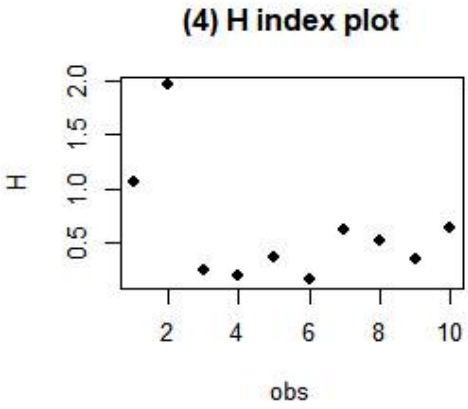
5) 잠재성-잔차플롯



6) Cook의 거리에 대한 인덱스플롯



7) Hadi의 영향력 측도에 대한 인덱스 플롯



4.10 표 4.7의 데이터에 대하여 비정상적인 관측개체를 식별하여라.

[표 4.7] 연습문제 4.10에 대한 데이터

번호	Y	X	번호	Y	X
1	8.11	0	7	9.60	19
2	11.00	5	8	10.30	20
3	8.20	15	9	11.30	21
4	8.30	16	10	11.40	22
5	9.40	17	11	12.20	23
6	9.30	18	12	12.90	24

Solve) 표준화잔차는 특이값의 존재를 평가하기 위해 가치 있는 정보를 제공한다.
하지만 높은 지레점은 작은 잔차를 가지는 경향이 있으므로, 비정상적인 관측개체를 찾기 위해서, 표준화잔차와 지레점을 함께 탐색하여 분석을 수행하는 것이 좋다.

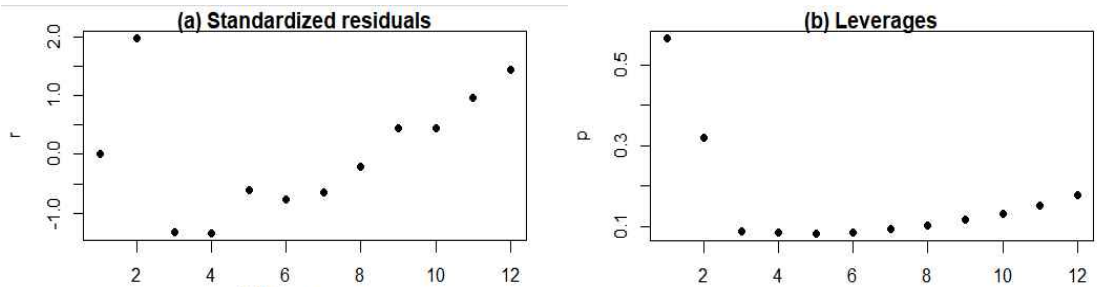
R코드 및 결과)

```
> data<-read.table("4.10.txt")
> data      #연습문제 4.10의 데이터
        Y  X
1    8.11  0
2   11.00  5
3    8.20 15
4    8.30 16
5    9.40 17
6    9.30 18
7    9.60 19
8   10.30 20
9   11.30 21
10  11.40 22
11  12.20 23
12  12.90 24
> a<-lm(Y~X, data = data)
> a

Call:
lm(formula = Y ~ X, data = data)

Coefficients:
(Intercept)            X
     8.1097       0.1235
```

```
> rstandard(a)  #(내적)표준화잔차
          1           2           3           4           5           6
          7           8
 0.0003288802  1.9691991445 -1.3186512619 -1.3332624346 -0.6037519783
-0.7718138773 -0.6420430380 -0.2105526464
          9          10          11          12
 0.4541716282  0.4405284070  0.9711373757  1.4391330972
> hatvalues(a)  #제레값
          1           2           3           4           5           6           7
          8           9          10
0.56502890  0.31936416  0.08815029  0.08410405  0.08352601  0.08641618  0.09277457
0.10260116 0.11589595 0.13265896
         11          12
0.15289017 0.17658960
> plot(rstandard(a), xlab="obs", ylab="r", main="(a) Standardized residuals", pch=16) #(내적)표준화잔차의 인덱스플롯
> plot(hatvalues(a), xlab="obs", ylab="p", main="(b) Leverages", pch=16) #지레값의 인덱스플롯
```



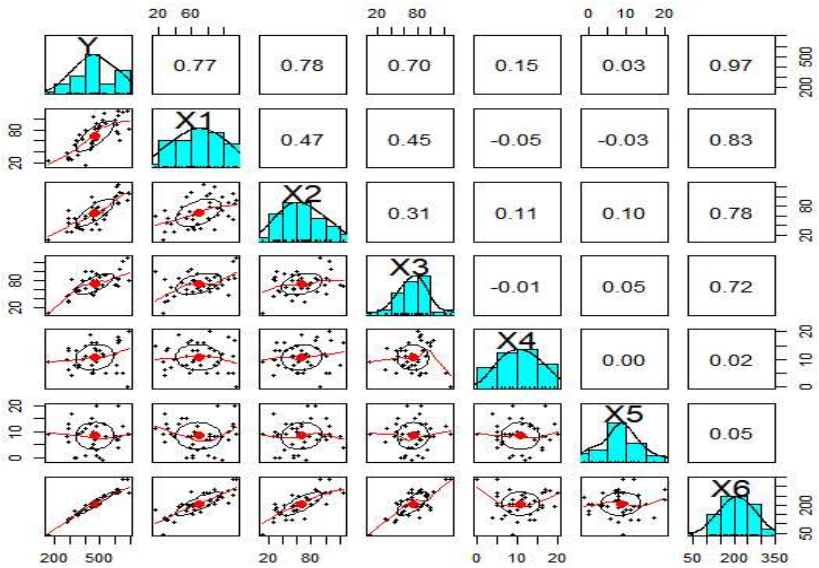
위의 (a)(내적)표준화잔차의 인덱스플롯에서 한 개의 특이점(2번)만 확인할 수 있다.
그러나 (b)지레값의 인덱스플롯에서는 두 개의 특이점(1, 2번)을 확인할 수 있다.
여기서 1번 관측개체는 (a)에서 작은 표준화잔차를 가짐에도 불구하고, (b)에서 큰 지레값을 가져 높은 지레점임을 알 수 있다.
이러한 결론은 표준화잔차에만 근거하여 분석을 수행할 경우, 비정상적인 관측개체(높은 지레점)가 존재하나 그것을 검출하지 못하는 가면문제가 발생할 수 있다는 것을 보여준다.

4.12 표 4.8의 데이터를 고려하자. 이것은 반응변수 Y와 여섯 개의 예측변수로 구성되어 있으며, 데이터는 이 책의 웹사이트에서도 구할 수 있다. Y와 여섯 개의 X-변수를 관계시키는 선형모형을 고려하자.

(a) 최소제곱의 가정이 위반되는 것으로 보이는 것은 무엇인가?(있다면)
(b) $r_i, C_i, DFITS_i, H_i$를 계산하여라.
(c) $r_i, C_i, DFITS_i, H_i$의 인덱스플롯과 잠재성-잔차플롯을 작성하여라.
(d) 데이터에 있는 비정상적인 관측개체를 식별하고 유형(특이값, 높은 지레점. 영향력 있는 개체 등)에 따라 분류하여라.

Solve)
(a) R코드 및 결과

```
> data2<-read.table("4.8.txt", header=TRUE)
> model<-lm(Y~X1+X2+X3+X4+X5+X6, data=data2)
> pairs.panels(data2)  #data2에 대한 산점도행렬
```



위의 그림은 표 4.8의 데이터에 대한 산점도 행렬이다.
X_1, X_6와 X_2, X_6의 상관계수가 각각 0.83 0.78이므로 다중공선성의 문제가 있다고 볼 수 있다 이는 모형의 형태에 대한 가정(선형성)과 예측변수들에 대한 가정에 문제를 일으킨다.
따라서 예측변수 $X_1, X_2 \cdots X_6$은 선형 독립이 아니므로, 최소제곱 해의 유일성이 보장되지 않는다. 그러므로 최소제곱의 가정이 위반된다고 볼 수 있다.

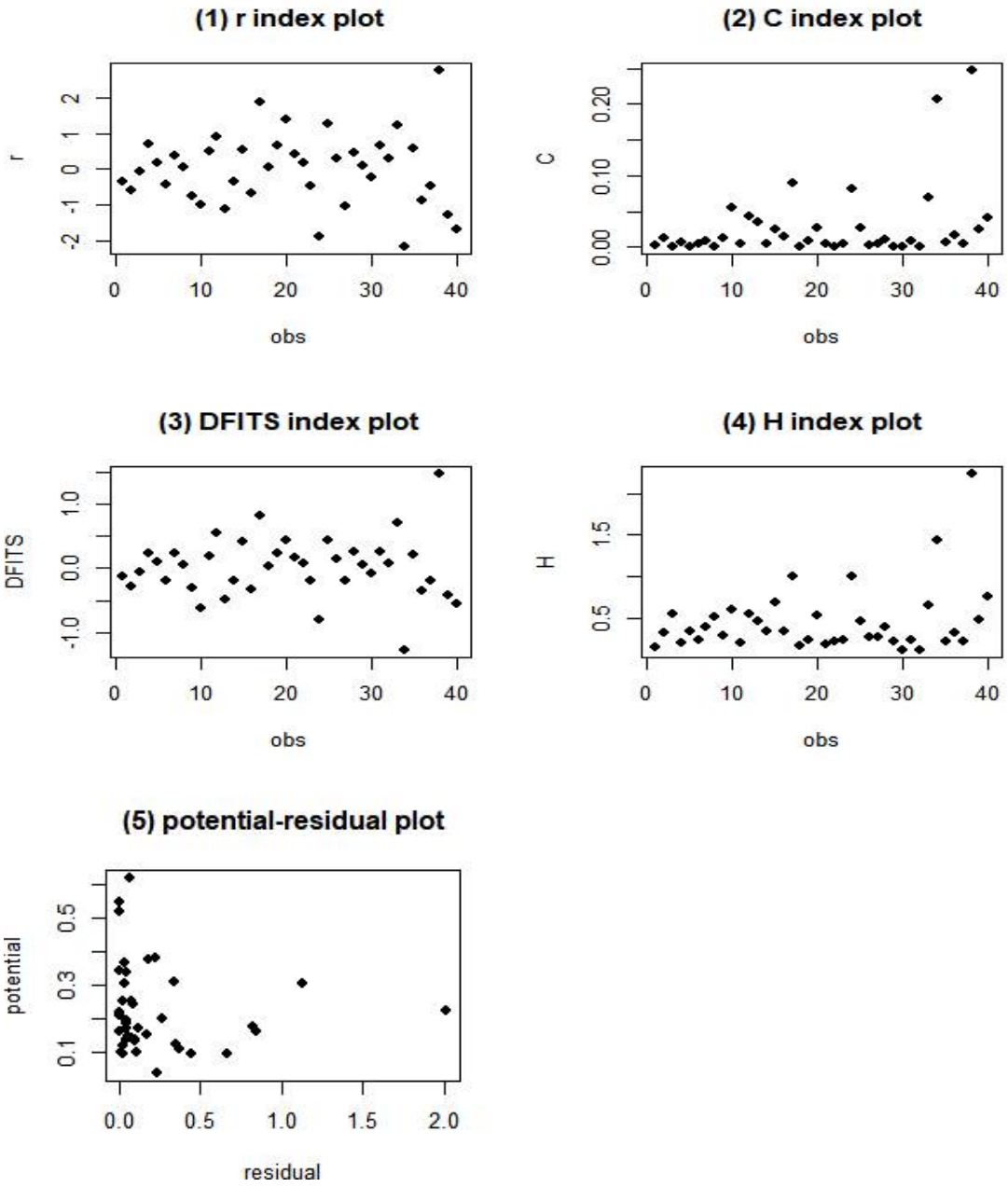
(b) R코드 및 결과

```
> data2<-read.table("4.8.txt", header=TRUE)
> model<-lm(Y~X1+X2+X3+X4+X5+X6, data=data2)
> cbind(rstandard(model), cooks.distance(model), dffits(model), ols_hadi(model)$hadi
) #1열 : r_i, 2열 : C_i, 3열 : DFIT_i, 4열 : H_i
```

	[,1]	[,2]	[,3]	[,4]
1	-0.35129222	0.0021423132	-0.12081534	0.1477836
2	-0.58952638	0.0124256541	-0.29196157	0.3246183
3	-0.07792919	0.0004744901	-0.05675715	0.5482097
4	0.71249421	0.0073966729	0.22581440	0.2112007
5	0.16181961	0.0012798353	0.09324305	0.3476867
6	-0.43825841	0.0053752442	-0.19157273	0.2368421
7	0.39778121	0.0082838730	0.23769905	0.4001559
8	0.07611439	0.0004277140	0.05388674	0.5180237
9	-0.75045785	0.0136865641	-0.30743417	0.2913460
10	-1.00173894	0.0545949061	-0.61822805	0.5984914
11	0.48758659	0.0049839161	0.18459607	0.1974943
12	0.90155776	0.0437099317	0.55153429	0.5519908
13	-1.10326197	0.0344324867	-0.49262000	0.4644129
14	-0.36310949	0.0057639955	-0.19819751	0.3340713
15	0.54165867	0.0259474755	0.42155459	0.6816516
16	-0.65044607	0.0147613086	-0.31858948	0.3349093
17	1.87655670	0.0897910089	0.82601415	0.9998374
18	0.07800872	0.0001423804	0.03109082	0.1650715
19	0.66966062	0.0087059351	0.24476305	0.2321717
20	1.40132735	0.0264029898	0.43653163	0.5346222
21	0.43815173	0.0038515912	0.16216386	0.1813706
22	0.16345503	0.0008321183	0.07518573	0.2236862
23	-0.46724397	0.0057956667	-0.19900321	0.2323987
24	-1.89358118	0.0821860718	-0.79112205	0.9996118
25	1.25027988	0.0279799124	0.44650674	0.4714534
26	0.30855896	0.0034466629	0.15317710	0.2736505
27	-1.03274658	0.0056995398	-0.19994985	0.2709231
28	0.46605746	0.0105317033	0.26825656	0.3857069
29	0.11520419	0.0003970040	0.05192205	0.2122064
30	-0.22949685	0.0007442923	-0.07113538	0.1101093
31	0.67217790	0.0089960124	0.24882024	0.2363803
32	0.28359222	0.0011126334	0.08701076	0.1139392
33	1.24624608	0.0689205737	0.70066389	0.6523500
34	-2.17595912	0.2072157606	-1.28147269	1.4346220
35	0.58516174	0.0070426632	0.21978621	0.2172718
36	-0.88199971	0.0168280034	-0.34202948	0.3198867
37	-0.47333320	0.0054539030	-0.19306344	0.2182026
38	2.77404672	0.2475472468	1.48031031	2.2412473
39	-1.28477977	0.0255334560	-0.42713445	0.4749701
40	-1.69798312	0.0403274572	-0.54767086	0.7623625

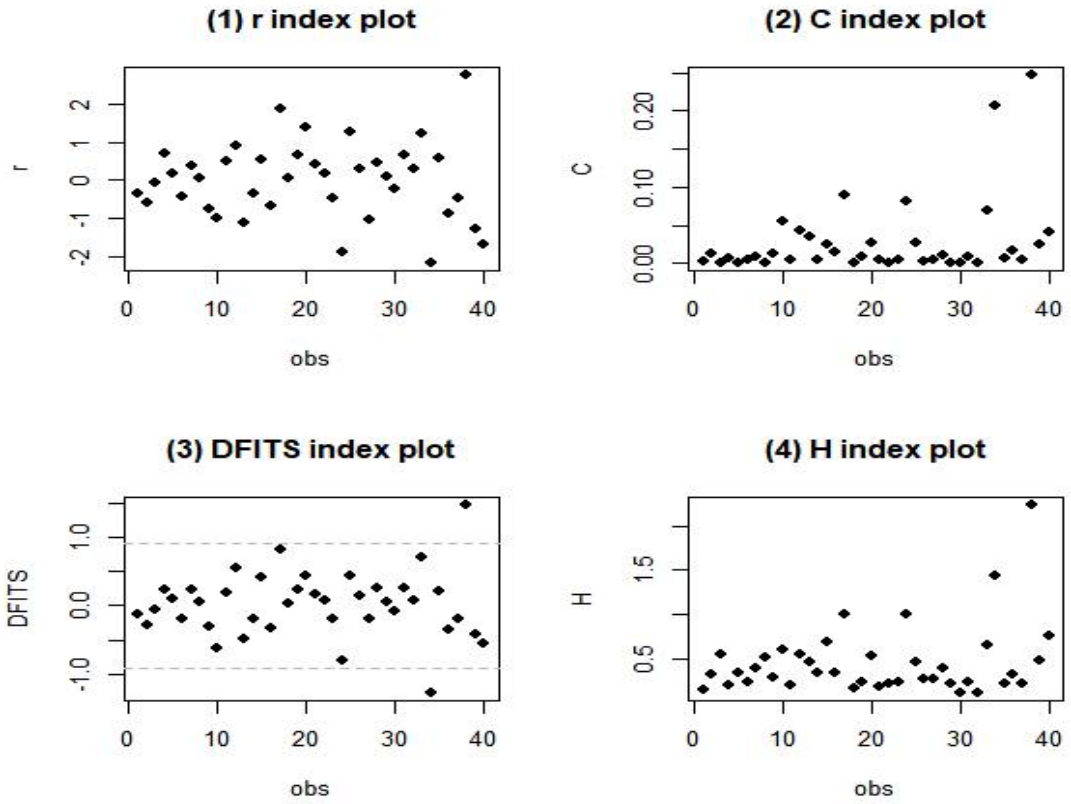
(c) R코드 및 결과

```
> par(mfrow=c(2,2))
> plot(rstandard(model), xlab="obs", ylab="r", main="(1) r index plot", pch=16)
> plot(cooks.distance(model), xlab="obs", ylab="C", main="(2) C index plot", pch=16)
> plot(dffits(model), xlab="obs", ylab="DFITS", main="(3) DFITS index plot", pch=16)
> plot(ols_hadi(model)$hadi, xlab="obs", ylab="H", main="(4) H index plot", pch=16)
> plot(ols_hadi(model)$residual, ols_hadi(model)$potential, xlab="residual", ylab="po
tential",main="(5) potential-residual plot", pch=16)
```



위 (1)번부터 (5)번까지 순서대로 r_i, C_i, $DFITS_i$, H_i의 인덱스플롯, 잠재성-잔차플롯이다.

(d) (c)의 (1)~(5)번 플롯을 이용한 비정상적인 관측개체의 분류는 다음과 같다.

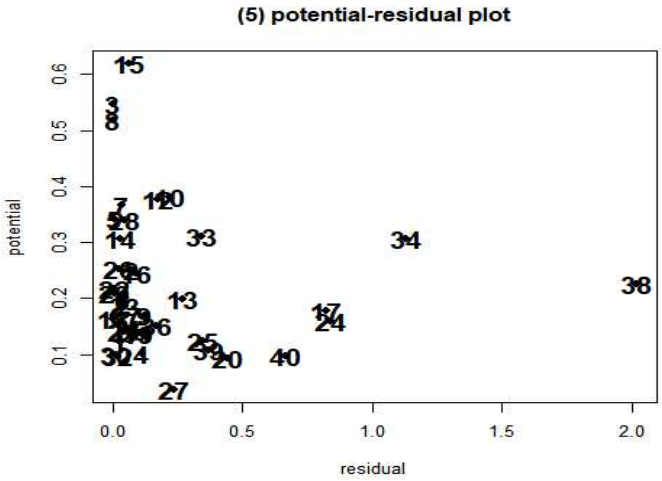


(1)번 플롯에서 17, 24, 34, 38번이 대략적으로 특이값을 가질 것으로 생각된다.
하지만 (1)번 플롯만으로 특이값을 고려하는 것은 섣부른 판단일 수 있으므로, (2)~(5)번 플롯을 확인한 뒤에 영향력과 특이값을 판단하여 분류해보도록 하겠다.
(2)번 플롯을 보면 34, 38번이 명확하게 영향력이 있고, 17, 24, 33번이 영향력 있다고 판단될 수 있다. 이어서 (3)번 플롯에선 $2\sqrt{(p+1)/(n-p-1)}=0.92$보다 큰 $|DFITS|$값인 34, 38번과 0.92에 근사한 17, 24, 33번이 영향력이 있다고 판단될 수 있다.
마지막으로 (4)번에서는 34, 38번을 명확하게 영향력 있는 개체라 할 수 있고, 3, 8, 15, 17, 24, 33, 40번 또한 영향력이 있다고 판단 될 수 있다,
정리해 보면 각 그래프에서 영향력 있다고 판단되는 개체들은 다음과 같다.

(2) C_i의 인덱스 플롯 : 17, 24, 33, 34, 38
(3) $DFITS_i$의 인덱스 플롯 : 17, 24, 33, 34, 38
(4) H_i의 인덱스 플롯 : 3, 8, 15, 17, 24, 33, 34, 38, 40

(2), (3)에는 (4)에서 영향력이 있다고 판단한 3, 8, 15, 40이 유의하지 않다고 판단되는데, (2) (3) 두 측도들은 지레값이 높다고 하더라도 잔차가 작으면 그 둘의 곱이 낮아져 낮은 값을 주기 때문이다. 이는 승법적 함수와 가법적 함수의 특징으로도 설명할 수 있을 것이라 생각한다
마지막으로 위의 측도에서 영향력이 있다고 판단된 3, 8, 15, 17, 24, 33, 34, 38, 40번의 개체를 위주로, (5) 잠재성-잔차 플롯을 확인해 보겠다.

다음은 (5) 잠재성-잔차플롯이다.



3, 8, 15번의 관측개체들은 잠재성이 높으므로(지레값이 크므로), 높은 지레점으로 판단된다. 또한 이 점들은 잔차가 작아 (2)cook's distance, (3)DFITS의 측도에서 영향력이 유의하다고 판단될 수 없었고, (4)Hadi의 측도에서도 눈에 띄게 높은 영향력을 가졌다고 판단할 수 없으므로 3, 8, 15번의 관측개체들은 영향력이 없는 높은 지레점으로 판단될 수 있다.

다음으로 34, 38번의 관측개체는 앞서 확인한 3, 8, 15번의 관측개체들 보다 높은 잠재성은 아니지만, 다른 관측개체들에 비해 비교적 높은 잠재성을 나타냄을 알 수 있다. 또한 잔차가 크기 때문에 (2)cook's distance, (3)DFITS, (4)Hadi의 측도에서 확연하게 높은 값을 가지며, 영향력이 있다고 할 수 있는 근거가 된다. 따라서 34, 38번의 관측개체는 가장 영향력이 있고, 특이점(지레값이 다른 관측개체들에 비해 명확히 크지 않고, (1)r_i의 인덱스 플롯에서 값이 2보다 크므로)으로 판단될 수 있다.

마지막으로 (4)에서만 영향력 있는 개체로 판단된 17, 24, 33, 40번 관측개체의 경우, (2), (3), (5)을 함께 고려하였을 때 명확하게 영향력을 가지고 있다고 하기 애매하고, 마찬가지로 특이점이나 높은 지레값이라고 하기 애매한 값을 가지므로, 정상적인 관측개체로 판단하는 것이 좋을 것 같다.

결과적으로 비정상적인 개체를 분류하면 다음과 같다.
1) 높은 지레점 : 3, 8, 15
2) 특이값 : 34, 38
3) 영향력 있는 개체 : 34, 38

5.3 5.6절에 제시된 내용들을 이용하여 표 5.11에 있는 스키 판매액 데이터에 대한 분석을 수행하여라.

Solve)

```
> Ski_Sales<-read.table("5.11.txt", header=TRUE)
> head(Ski_Sales)     #표 5.11 스키 판매액 데이터
     Date Sales PDI
1  Q1/64  37.0 109
2  Q2/64  33.5 115
3  Q3/64  30.8 113
4  Q4/64  37.9 116
5  Q1/65  37.4 118
6  Q2/65  31.6 120
```

표 5.11의 스키 판매액 데이터의 분석에 고려되는 축소모형과 완전모형은 다음과 같다.

$$RM: S_t = \beta_0 + \beta_1 PDI_t + \epsilon_t$$
$$FM: S_t = \beta_0 + \beta_1 PDI_t + \gamma_1 z_{t1} + \gamma_2 z_{t1} + \gamma_3 z_{t3} + \delta_1 z_{t1} PDI_t + \delta_2 z_{t2} PDI_t + \delta_3 z_{t3} PDI_t + \epsilon_t$$

단, S_t와 PDI_t는 각각 기간 t의 매출액(Sales)과 개인 가처분소득액(PDI)이다.
또한 $z_{ti}, i = 1, 2, 3$는 분기/년도(Date)에 대한 가변수로 다음과 같이 정의된다.

$$z_{t1} = \begin{cases} 1, & t\text{번째 기간이 } 1/4\text{분기인 경우} \\ 0, & \text{그외의 경우} \end{cases}$$

$$z_{t2} = \begin{cases} 1, & t\text{번째 기간이 } 2/4\text{분기인 경우} \\ 0, & \text{그외의 경우} \end{cases}$$

$$z_{t3} = \begin{cases} 1, & t\text{번째 기간이 } 3/4\text{분기인 경우} \\ 0, & \text{그외의 경우} \end{cases}$$

위의 모형을 결정하기 위해서 관심의 대상이 되는 가설은 아래와 같다.

$$H_0 : \gamma_1 = \gamma_2 = \gamma_3 = \delta_1 = \delta_2 = \delta_3 = 0$$
$$H_1 : not\ H_0$$

가설을 검정하기 위한 과정은 다음과 같다.

```
> Ski_Sales<-read.table("5.11.txt", header=TRUE)
> Z1 <- ifelse(substr(Ski_Sales$Date, 1, 2)==Q1, 1, 0)
> Z2 <- ifelse(substr(Ski_Sales$Date, 1, 2)==Q2, 1, 0)
> Z3 <- ifelse(substr(Ski_Sales$Date, 1, 2)==Q3, 1, 0)
```

위의 코드는 분기(Q1, Q2, Q3, Q4)를 반영하기 위해 3개의 가변수를 정의하는 과정이다.
다음은 축소모형과 완전모형의 회귀분석 결과이다.

```
> RM<-lm(Sales~PDI, data=Ski_Sales)
> summary(RM)

Call:
lm(formula = Sales ~ PDI, data = Ski_Sales)

Residuals:
    Min      1Q  Median      3Q     Max
-5.229 -2.686   0.554   2.728   4.454

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.39215    2.53942    4.88 1.93e-05 ***
PDI          0.19791    0.01602   12.35 7.09e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.019 on 38 degrees of freedom
Multiple R-squared:  0.8006,    Adjusted R-squared:  0.7953
F-statistic: 152.5 on 1 and 38 DF,  p-value: 7.089e-15

> FM<-lm(Sales~ PDI + Z1+Z2+Z3 + PDI*Z1 +PDI*Z2 + PDI*Z3, data=Ski_Sales)
> summary(FM)

Call:
lm(formula = Sales ~ PDI + Z1 + Z2 + Z3 + PDI * Z1 + PDI * Z2 +
    PDI * Z3, data = Ski_Sales)

Residuals:
      Min        1Q    Median        3Q       Max
-2.48116 -0.88688  0.02911  0.68862  2.67909

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.880359   2.094256   7.105 4.63e-08 ***
PDI          0.198363   0.012948  15.320 2.74e-16 ***
Z1           0.053930   2.894707   0.019   0.9853
Z2          -5.601102   2.935307  -1.908   0.0654 .
Z3          -5.105100   2.931804  -1.741   0.0912 .
PDI:Z1       0.001939   0.018291   0.106   0.9163
PDI:Z2       0.002029   0.018374   0.110   0.9127
```

PDI:Z3 -0.001202 0.018245 -0.066 0.9479

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.214 on 32 degrees of freedom
Multiple R-squared: 0.9728, Adjusted R-squared: 0.9669
F-statistic: 163.8 on 7 and 32 DF, p-value: < 2.2e-16

```
>Fval<-((sum((RM$residuals)^2)-sum((FM$residuals)^2))/(8-2))/(sum((FM$residuals)^2)/(40-7-1))
> Fval
[1] 33.82882
> qf(0.95, 6, 32)
[1] 2.39908
```

축소모형(RM)과 완전모형(FM)의 회귀분석 결과를 통해, F검정통계량이 33.82882임을 구할 수 있었다. 그런데 F값이 유의수준 5%의 임계치 2.3998보다 큰 값을 가지므로, 귀무가설을 기각할 수 있다.
이러한 결과로 인해, 분기별(Data) 집단에 따른 매출액(Sales)과 가처분소득(PDI)의 회귀관계가 같지 않다고 할 근거가 충분하다.

다음은 분기별 추정된 회귀직선이다.

Date=Q1(1/4분기)일 때, $S_t = 14.934 + 0.2PDI_t$
Date=Q2(2/4분기)일 때, $S_t = 9.279 + 0.2PDI_t$
Date=Q3(3/4분기)일 때, $S_t = 9.775 + 0.197PDI_t$
Date=Q4(4/4분기)일 때, $S_t = 14.88 + 0.198PDI_t$

5.4 5.7절에 제시된 내용들을 이용하여 표 5.12, 5.13, 5.14에 있는 교육비 지출 데이터에 대한 분석을 수행하여라

Solve)
문제 5.4의 교육비 지출 데이터의 분석에 고려되는 축소모형과 완전모형은 다음과 같다.

$$RM: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$
$$FM: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \gamma_1 T_1 + \gamma_2 T_2 + \delta_1 T_1 X_1 + \delta_2 T_1 X_2 + \delta_3 T_1 X_3 + \alpha_1 T_2 X_1 + \alpha_2 T_2 X_2 + \alpha_3 T_2 X_3 + \epsilon$$

단, Y는 공공교육에 사용되는 개인의 교육비 지출이고 X_1, X_2, X_3는 각각 개인소득, 1,000명당 18세 미만의 인구수, 1,000명당 도심 거주 인구수이다.
또한 $T_{ij}, j = 1, 2$는 3개년(1960년, 1970년, 1975년)에 대한 가변수로 다음과 같이 정의된다.

$$T_{i1} = \begin{cases} 1, & i\text{번째 관측값이 1960년에 해당하는 경우} \\ 0, & \text{그외의 경우} \end{cases}$$

$$T_{i2} = \begin{cases} 1, & i\text{번째 관측값이 1970년에 해당하는 경우} \\ 0, & \text{그외의 경우} \end{cases}$$

위의 모형을 결정하기 위해서 관심의 대상이 되는 가설은 아래와 같다.

$$H_0 : \gamma_1 = \gamma_2 = \gamma_3 = \delta_1 = \delta_2 = \delta_3 = \alpha_1 = \alpha_2 = \alpha_3 = 0$$
$$H_1 : not\ H_0$$

가설을 검정하기 위한 과정은 다음과 같다.

```
> data_1960<-read.table("5.12.txt", header=TRUE)
> data_1970<-read.table("5.13.txt", header=TRUE)
> data_1975<-read.table("5.14.txt", header=TRUE)
> data<-rbind(cbind(data_1960,YEAR=1960), cbind(data_1970, YEAR=1970), cbind(dat
a_1975, YEAR=1975))
> head(data)
  STATE   Y    X1  X2  X3 Region YEAR
1    ME 61 1704 388 399        1 1960
2    NH 68 1885 372 598        1 1960
3    VT 72 1745 397 370        1 1960
4    MA 72 2394 358 868        1 1960
5    RI 62 1966 357 899        1 1960
6    CT 91 2817 362 690        1 1960
```

위의 과정은 분석을 위해 표 5.12, 5.13, 5.14의 데이터에 YEAR 변수를 추가하여 각각 1960년, 1970년, 1975년의 데이터임을 나타낸 후, rbind함수를 통해 하나의 데이터로 결합하였다. 이것은 다음에 이어질 가변수를 정의하고, 완전모형의 회귀분석을 위한 필수적인 과정이다.

```
> T1<-ifelse(data$YEAR == 1960, 1, 0)
> T2<-ifelse(data$YEAR == 1970, 1, 0)
```

위의 코드는 3개년(1960년, 1970년, 1975년)을 반영하기 위해 2개의 가변수를 정의하는 과정이다.
다음은 축소모형과 완전모형의 회귀분석 결과이다.

```
> RM<-lm(Y~X1+X2+X3, data=data)
> summary(RM)

Call:
lm(formula = Y ~ X1 + X2 + X3, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-87.890 -18.531  -3.049  16.798 153.381

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.419e+02  4.059e+01  -3.496 0.000627 ***
X1           8.021e-02  3.448e-03  23.261  < 2e-16 ***
X2           2.979e-01  8.431e-02   3.533 0.000550 ***
X3          -6.340e-02  2.209e-02  -2.870 0.004716 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 35.67 on 146 degrees of freedom
Multiple R-squared:  0.8582,    Adjusted R-squared:  0.8553
F-statistic: 294.7 on 3 and 146 DF,  p-value: < 2.2e-16
```

```
> FM<-lm(Y~ X1+X2+X3 + T1+T2 + T1*X1+T1*X2+T1*X3 + T2*X1+T2*X2+T2*X3,
data=data)
> summary(FM)

Call:
lm(formula = Y ~ X1 + X2 + X3 + T1 + T2 + T1 * X1 + T1 * X2 +
    T1 * X3 + T2 * X1 + T2 * X2 + T2 * X3, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-84.878 -16.644  -2.204  16.312  99.243
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-5.566e+02	8.981e+01	-6.197	6.23e-09	***
X1	7.239e-02	8.459e-03	8.558	1.99e-14	***
X2	1.552e+00	2.294e-01	6.766	3.47e-10	***
X3	-4.269e-03	3.747e-02	-0.114	0.9095	
T1	5.452e+02	1.050e+02	5.190	7.36e-07	***
T2	2.674e+02	1.154e+02	2.318	0.0219	*
X1:T1	-2.745e-02	1.671e-02	-1.643	0.1027	
X2:T1	-1.486e+00	2.471e-01	-6.013	1.54e-08	***
X3:T1	-2.468e-02	5.215e-02	-0.473	0.6367	
X1:T2	8.501e-03	1.338e-02	0.635	0.5262	
X2:T2	-7.336e-01	2.896e-01	-2.533	0.0124	*
X3:T2	-9.950e-02	5.362e-02	-1.856	0.0656	.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.51 on 138 degrees of freedom
Multiple R-squared: 0.9083, Adjusted R-squared: 0.901
F-statistic: 124.3 on 11 and 138 DF, p-value: < 2.2e-16

```
>Fval<-((sum((RM$residuals)^2)-sum((FM$residuals)^2))/(11-3))/(sum((FM$residuals)^2)
/(150-11-1))
> Fval
[1] 9.422465
> qf(0.95, 8, 138)
[1] 2.006109
```

축소모형(RM)과 완전모형(FM)의 회귀분석 결과를 통해, F검정통계량이 9.422465임을 구할 수 있었다. 그런데 F값이 유의수준 5%의 임계치 2.006109보다 큰 값을 가지므로, 귀무가설을 기각할 수 있다
따라서 3개년(1960년, 1970년, 1975년) 모두 똑같은 변수들로 규정되는 회귀관계가 같지 않다고 할 충분한 근거를 갖는다.

다음은 1960년, 1970년, 1975년의 각각 추정된 회귀직선이다.

YEAR=1960(1960년)일 때, $Y = -11.405 + 0.045X_1 + 0.066X_2 - 0.029X_3$
YEAR=1970(1970년)일 때, $Y = -289.179 + 0.081X_1 + 0.818X_2 - 0.103X_3$
YEAR=1975(1975년)일 때, $Y = -556.568 + 0.072X_1 + 1.552X_2 - 0.004X_3$

5.7 더 많은 옥수수를 생산할 수 있는지를 알아보기 위하여 세 종류의 비료를 시험하였다. 40개의 유사한 토지 구획이 사용 가능하였다. 40개의 구획은 4개의 집단으로 랜덤하게 구분되었고, 각 집단에 10개씩의 구획이 할당되었다. 집단 4에 있는 옥수수에는 비료가 주어지지 않았으며, 이는 대조집단의 역할을 할 것이다. 표 5.17은 40개의 각 구획에 대한 옥수수 생산량 y_{ij}를 제시한다.

(a) 각 비료에 대하여 하나씩 새 개의 지시변수 F_1, F_2, F_3을 만들어라.
(b) 모형 $y_{ij} = \mu_0 + \mu_1 F_{i1} + \mu_2 F_{i2} + \mu_3 F_{i3} + \epsilon_{ij}$를 적합하여라.
(c) 평균적으로 세 종류의 비료 중 어느 것도 옥수수 생산량에 영향을 미치지 않는다는 가설을 검정하여라. 검정될 가설, 검정법, 5% 유의수준 하에서의 결론을 구체화하여라
(d) 평균적으로 세 종류의 비료가 옥수수 생산량에 미치는 영향이 같으나 대조 집단의 그것과는 다르다는 가설을 검정하여라. 검정될 가설, 검정법, 5% 유의수준 하에서의 결론을 구체화하여라.
(e) 세 종류의 비료 중 어느 것이 옥수수 생산량에 가장 큰 영향을 미치는가?

Solve)
(a) 각 비료에 대한 세 개의 지시변수 F_1, F_2, F_3는 다음과 같다.

$$F_{i1} = \begin{cases} 1, & i\text{번째 비료가 비료1인 경우} \\ 0, & \text{그외의 경우} \end{cases}$$

$$F_{i2} = \begin{cases} 1, & i\text{번째 비료가 비료2인 경우} \\ 0, & \text{그외의 경우} \end{cases}$$

$$F_{i3} = \begin{cases} 1, & i\text{번째 비료가 비료3인 경우} \\ 0, & \text{그외의 경우} \end{cases}$$

(b) R코드 및 결과

```
> data<-read.table("5.17.txt", header=TRUE)
> head(data)
  Yield Fertilizer
1    31          1
2    34          1
3    34          1
4    34          1
5    43          1
6    35          1
> F1 <- ifelse(data$Fertilizer==1, 1, 0)
> F2 <- ifelse(data$Fertilizer==2, 1, 0)
> F3 <- ifelse(data$Fertilizer==3, 1, 0)
> model<-lm(Yield~F1+F2+F3, data=data)
> model
```

Call:
lm(formula = Yield ~ F1 + F2 + F3, data = data)

Coefficients:

(Intercept)	F1	F2	F3
29.8	6.8	0.1	5.1

따라서 적합된 모형은 $Y = 29.8 + 6.8F_1 + 0.1F_2 + 5.1F_3$이다.

(c) 평균적으로 세 종류의 비료 중 어느 것도 옥수수 생산량에 영향을 미치지 않는다는 가설은 다음과 같다.

$H_0 : \mu_1 = \mu_2 = \mu_3 = 0$
$H_1 : not\ H_0$

R코드 및 결과

```
> model<-lm(Yield~F1+F2+F3, data=data)
> model1=lm(Yield~1, data=data)
> anova(model,model1)
```

Analysis of Variance Table

Model 1: Yield ~ F1 + F2 + F3
Model 2: Yield ~ 1

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	36	845.8				
2	39	1208.4	-3	-362.6	5.1445	0.004605 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

F값의 p-value는 0.004605로 유의수준 5% 귀무가설을 기각할 근거가 충분하다. 따라서 3종류의 비료 중 어느 것도 옥수수에 영향을 미친다고는 하지 못한다.

(d) 세 종류의 비료가 옥수수 생산량에 미치는 영향이 같으므로, 지시변수는 다음과 같이 다시 정의 할 수 있다.

$$F_{i1} = \begin{cases} 1, \text{비료처리를 한 관측개체의 경우} \\ 0, \text{그외의 경우} \end{cases}$$

이를 토대로 대조집단과 다르다는 가설을 검정하기 위한 H_0, H_1은 다음과 같다.

$H_0 : \mu' = 0$
$H_1 : \mu' \neq 0$

가설을 검정하기 위한 과정은 다음과 같다.

```
> F1 <- ifelse(data$Fertilizer==1 | data$Fertilizer==2 | data$Fertilizer==3, 1, 0)
> F1
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0
0 0
> model1<-lm(Yield~F1, data=data)
> summary(model1)

Call:
lm(formula = Yield ~ F1, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-12.80  -3.05    0.20    3.20   11.20

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   29.800      1.692  17.608   <2e-16 ***
F1             4.000      1.954   2.047   0.0476 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.352 on 38 degrees of freedom
Multiple R-squared:  0.0993,    Adjusted R-squared:  0.0756
F-statistic:  4.19 on 1 and 38 DF,  p-value: 0.04763
```

F1의 회귀계수, 즉 μ'에 대해 p-value가 0.0476으로 작아 유의하다는 결론을 얻을 수 있다. 따라서 귀무가설을 기각할 근거가 충분하여, 세 종류의 비료가 옥수수 생산량에 미치는 영향은 같으나 대조집단의 그것과는 다르다고 할 수 있다.

(e) 모형 $y_{ij} = \mu_0 + \mu_1 F_{i1} + \mu_2 F_{i2} + \mu_3 F_{i3} + \epsilon_{ij}$의 회귀분석 결과이다.

```
> summary(model)

Call:
lm(formula = Yield ~ F1 + F2 + F3, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.800 -2.825 -0.600   3.125  10.200

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   29.800      1.533  19.442   <2e-16 ***
F1             6.800      2.168   3.137   0.0034 **
F2             0.100      2.168   0.046   0.9635
F3             5.100      2.168   2.353   0.0242 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.847 on 36 degrees of freedom
Multiple R-squared:  0.3001,    Adjusted R-squared:  0.2417
F-statistic: 5.144 on 3 and 36 DF,  p-value: 0.004605
```

비료1, 비료2, 비료3에 대한 회귀계수의 p-value를 확인할 수 있다. 비료1과 비료3의 회귀계수는 유의한 반면, 비료2의 회귀계수는 유의하지 않음을 알 수 있다. 따라서 비료1과 비료3이 비료에 미치는 영향만 확인하는 것이 좋다고 생각한다.
위의 비료1과 비료3의 회귀계수가 각각 6.8, 5.1이므로, 비료1이 옥수수 생산량에 가장 큰 영향력을 미친다고 할 수 있다.