9.4 표 5.19의 대통령 선거 데이터를 참조하여 V와 모든 예측변수(선거 연도를 나타내는 시간 경향 변수도 포함)에 더하여 2차 3차 교호작용 항을 가능한 많이 포함하는 회귀모형을 고려하자.

(a) 이러한 데이터를 적합하기 위한 선형회귀모형에서 최대 항의 수는 얼마인가?
[힌트 : 데이터에서 관측개체의 수를 생각하여 보아라.]
(b) 위의 모형에 대한 다중공선성 여부를 위해 각 예측변수들을 조사하여 보아라.
(상관계수 행렬, 상태수, **VIF**를 계산하여라.)
(c) 공선성을 포함하는 변수들의 집합을 찾아라. 다중공선성을 가지는 몇몇 변수들을 제거하여 다중공선성의 문제를 해결하여 보아라.
(d) V와 다중공선성이 없다고 판단되는 예측변수들에 대해 회귀모형을 적합시켜라.

Solve)
표 5.19의 변수들은 다음과 같다.

변수	정의
Year	선거 연도
V	(민주/공화) 양당투표에서 민주당의 득표율
I	지시변수. 현직자가 민주당 소속인 경우-1, 현직자가 공화당 소속인 경우=-1
D	범주형 변수, 민주당 현직자가 입후보한 경우=1, 공화당 현직자가 입후보한 경우=1, 그렇지 않으면 0
W	지시변수. 1920, 1944, 1948년 대하여는 1, 그렇지 않으면 0
G	선거 당해년의 첫 3분기에서 실질 1인당 GDP 성장률(%)
P	현 행정부의 첫 15분기에서 GDP 성장률의 절댓값(%)
N	현 행정부의 첫 15분기에서 실질 1인당 GDP 성장률이 3.2% 이상인 분기의 수

(a) 문제에서 예측변수에 더하여 2차 또는 3차 교호작용 항을 가능한 많이 포함하는 회귀모형을 고려하자고 하였으므로 만들 수 있는 항들은 다음과 같다.

1) 1차항
: Year, I, D, W, G, P, N
2) 2차항(2차 교호작용 항)
: Year$*I$, Year$*D$, Year$*W$, Year$*G$, Year$*P$, Year$*N$,
$I*D$, $I*W$, $I*G$, $I*P$, $I*N$,
$D*W$, $D*G$, $D*P$, $D*N$,
$W*G$, $W*P$, $W*N$,
$G*P$, $G*N$,
$P*N$
3) 3차항(3차 교호작용 항)
: Year$*I*D$, Year$*I*W$, Year$*I*G$, Year$*I*P$, Year$*I*N$,
Year$*D*W$, Year$*D*G$, Year$*D*P$, Year$*D*N$,
Year$*W*G$, Year$*W*P$, Year$*W*N$,
Year$*G*P$, Year$*G*N$,
Year$*P*N$,

$I*D*W$, $I*D*G$, $I*D*P$, $I*D*N$,
$I*W*G$, $I*W*P$, $I*D*N$,
$I*G*P$, $I*G*N$,
$I*P*N$,
$D*W*G$, $D*W*P$, $D*W*N$,
$D*G*P$, $D*G*N$,
$D*P*N$,
$W*G*P$, $W*G*N$,
$W*P*N$,
$G*P*N$

1차항(7개), 2차항(2차 교호작용 항)(21개), 3차항(3차 교호작용 항)(35개)을 모두 더해 63개의 항이 회귀모형에서 고려될 수 있다. 그런데 문제의 관측개체 수가 21개이므로, 회귀모형의 자유도 $n-p-1$을 고려하였을 때, p는 최대 19개의 항을 가질 수 있다.
따라서 기존의 예측변수(Year, I, D, W, G, P, N) 7개에 임의로 선택한 2, 3차 교호작용 항 12개를 더하여 p를 정할 수 있고, 이를 통해 회귀모형을 정할 수 있을 것이다.
기존의 예측변수를 포함하고, 임의로 교호작용 항을 선택하여 만든 회귀모형은 다음과 같다.

$$V = \beta_0 + \beta_1 Year + \beta_2 I + \beta_3 D + \beta_4 W + \beta_5 G + \beta_6 P + \beta_7 N + \beta_8 (I \cdot G) + \beta_9 (I \cdot P) + \beta_{10}(I \cdot N) + \beta_{11}(D \cdot G) + \beta_{12}(D \cdot P) + \beta_{13}(D \cdot N) + \beta_{14}(I \cdot G \cdot P) + \beta_{15}(I \cdot G \cdot N) + \beta_{16}(I \cdot P \cdot N) + \beta_{17}(D \cdot G \cdot P) + \beta_{18}(D \cdot G \cdot N) + \beta_{19}(D \cdot P \cdot N) + \epsilon$$

(b) 각 예측변수의 상관계수 행렬, 상태수, **VIF**는 다음과 같다.

```
> data<-read.table("5.19.txt", header=TRUE)
> head(data, 3)
  Year      V  I  D W       G      P  N
1 1916 0.5168  1  1 0   2.229  4.252  3
2 1920 0.3612  1  0 1 -11.463 16.535  5
3 1924 0.4176 -1 -1 0  -3.872  5.161 10

> cor(data[,-2])   #상관행렬
           Year          I           D          W          G           P          N
Year  1.0000000 -0.2046969 -0.11637147 -0.3146266  0.3085929 -0.18482213 -0.3161760
I    -0.2046969  1.0000000  0.81744307  0.3892495  0.1369564  0.11921266  0.2650860
D    -0.1163715  0.8174431  1.00000000  0.2876780  0.3230490 -0.07290826  0.2835083
W    -0.3146266  0.3892495  0.28767798  1.0000000 -0.2168366  0.64831150  0.2718636
G     0.3085929  0.1369564  0.32304903 -0.2168366  1.0000000 -0.58368979  0.2617113
P    -0.1848221  0.1192127 -0.07290826  0.6483115 -0.5836898  1.00000000 -0.1670507
N    -0.3161760  0.2650860  0.28350827  0.2718636  0.2617113 -0.16705075  1.0000000

> sqrt(max(eigen(cor(data[,-2]))$values)/eigen(cor(data[,-2]))$values)  #상태수
[1] 1.000000 1.080845 1.510115 1.839639 2.609045 3.603208 4.051070
```

```
> VIF(lm(V~Year+I+D+W+G+P+N, data=data)) #VIF
    Year        I        D        W        G        P        N 
1.432490 3.433255 3.620744 2.746286 2.095444 3.194597 1.561810 
```

(c) (b)에서 상관행렬과 상태지수, **VIF**를 알아보았다. 상태수와 각 예측변수의 **VIF**이 확연히 큰 값을 갖지 않는 것으로 보아, 강한 다중공선성을 보이는 변수는 없을 것으로 생각된다.
다만 상관행렬에서 변수 I와 D의 상관계수가 0.817인 점과 각각의 **VIF**가 다른 변수보다 상대적으로 높은 3.43, 3.62인 점을 고려하였을 때, 그 둘 사이에 어느 정도의 공선성이 존재한다고 할 수 있다. 변수 P 또한 W, G와 각각 상관계수 0.64, -0.58를 갖고, I, D와 비슷한 VIF를 가지는 것으로 보아, 공선성이 발견될 수 있을 것으로 생각된다.
따라서 변수 I와 P를 제거하여, 다중공선성의 문제가 해결되었는지 확인해 보도록 하자.
다음은 변수 I와 P를 제거한 예측변수들의 상관행렬과 상태지수 **VIF**이다.

```
> head(data[,-c(2, 3, 7)], 3)
  Year  D W       G  N
1 1916  1 0   2.229  3
2 1920  0 1 -11.463  5
3 1924 -1 0  -3.872 10
> cor(data[,-c(2, 3, 7)])
           Year          D          W          G          N
Year  1.0000000 -0.1163715 -0.3146266  0.3085929 -0.3161760
D    -0.1163715  1.0000000  0.2876780  0.3230490  0.2835083
W    -0.3146266  0.2876780  1.0000000 -0.2168366  0.2718636
G     0.3085929  0.3230490 -0.2168366  1.0000000  0.2617113
N    -0.3161760  0.2835083  0.2718636  0.2617113  1.0000000

> sqrt(max(eigen(cor(data[,-c(2, 3, 7)]))$values)/eigen(cor(data[,-c(2, 3, 7)]))$values) # 상태수
[1] 1.000000 1.102356 1.507900 1.749856 2.317262

> VIF(lm(V~Year+D+W+G+N, data=data)) #VIF
    Year        D        W        G        N 
1.408476 1.347613 1.361432 1.629309 1.425504 
```

위의 상태지수에서 상태수가 2.31로 줄었음을 확인할 수 있다. 또한 변수 I와 공선성이 있을 것으로 생각되었던 D의 **VIF**가 크게 줄었고, 마찬가지로 변수 P와 공선성이 있을 것으로 생각된 W, G의 **VIF** 또한 크게 감소했음을 알 수 있다.
위의 상태수와 **VIF**의 감소를 통해, 다중공선성의 문제를 감소시켰다고 볼 수 있을 것이다.

(d) (a)에서 고려되었던 회귀모형은 다음과 같다.

$$V=\beta_0+\beta_1 Year+\beta_2 I+\beta_3 D+\beta_4 W+\beta_5 G+\beta_6 P+\beta_7 N+\beta_8(I\cdot G)+\beta_9(I\cdot P)+\beta_{10}(I\cdot N)+\beta_{11}(D\cdot G)+\beta_{12}(D\cdot P)+\beta_{13}(D\cdot N)+\beta_{14}(I\cdot G\cdot P)+\beta_{15}(I\cdot G\cdot N)+\beta_{16}(I\cdot P\cdot N)+\beta_{17}(D\cdot G\cdot P)+\beta_{18}(D\cdot G\cdot N)+\beta_{19}(D\cdot P\cdot N)+\epsilon$$

그러나 (c)를 통해 변수 I와 P가 다중공선성의 문제를 일으키는 변수 집합임을 확인하였다. 따라서 (a)의 회귀모형에서 다중공선성이 없다고 판단되는 항들을 예측변수로 하는 새로운 회귀모형을 고려할 수 있다. 고려된 회귀모형은 다음과 같다.

$$V = \beta_0 + \beta_1 Year + \beta_2 D + \beta_3 W + \beta_4 G + \beta_5 N + \beta_6 (D \cdot G) + \beta_7 (D \cdot N) + \beta_8 (D \cdot G \cdot N) + \epsilon$$

다음은 회귀모형을 적합하는 과정이다.

```
> summary(lm(V~Year+D+W+G+N+D*G+D*N+D*G*N, data=data))

Call:
lm(formula = V ~ Year + D + W + G + N + D * G + D * N + D * G *
    N, data = data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.037052 -0.018275  0.003418  0.019619  0.035800

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.1619312  0.7689201  -1.511   0.1589
Year         0.0008382  0.0003904   2.147   0.0549 .
D           -0.0304768  0.0248061  -1.229   0.2449
W           -0.0183503  0.0265581  -0.691   0.5039
G            0.0114562  0.0063710   1.798   0.0996 .
N           -0.0019313  0.0040491  -0.477   0.6427
D:G          0.0185647  0.0061849   3.002   0.0120 *
D:N          0.0113932  0.0045114   2.525   0.0282 *
G:N         -0.0014928  0.0010274  -1.453   0.1742
D:G:N       -0.0007628  0.0009459  -0.806   0.4371
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03142 on 11 degrees of freedom
Multiple R-squared:  0.9036,	Adjusted R-squared:  0.8248
F-statistic: 11.46 on 9 and 11 DF,  p-value: 0.000208
```

따라서 적합된 회귀식은 다음과 같다.

$$V = -1.1619 + 0.0008\,Year - 0.0304D - 0.0183\,W + 0.0114G - 0.0019N + 0.0185(D \cdot G) + 0.0113(D \cdot N) - 0.0008(D \cdot G \cdot N) + \epsilon$$

9.5 표 5.9의 대통령 선거자료와 모형 (5.12)의 적합을 고려하자.

(a) 모형에 포함되는 예측변수들에 대한 다중공선성의 존재 여부를 살펴보아라.
(상관행렬과 상태지수, **VIF**를 계산하여라.)

(b) 다중공선성을 보이는 변수들의 그룹을 찾아내어라. 또한 특정 변수들을 제거함으로써 다중공선성의 문제를 해결하여 보여라.

(c) 다중공선성이 없는 예측변수들과 V의 회귀모형을 적합하여라.

(d) 연습문제 9.4에서의 결과와 비교하여라.

Solve)

(a) 모형 (5.12)는 다음과 같다.

$$V = \beta_0 + \beta_1 I + \alpha_1 D_1 + \alpha_2 D_2 + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \epsilon$$

여기서 D는 범주형 변수를 가변수로 변환한 것으로 다음과 같이 정의된다.

$D_1 = 1(\text{if } D = 1),\ D_1 = 0(\text{if } D \neq 1)$
$D_2 = 1(\text{if } D = -1),\ D_2 = 0(\text{if } D \neq -1)$

다음은 모형에 포함되는 예측변수들의 상관행렬과 상태지수, VIF를 구하는 과정이다.

```
> GI<-data$G*data$I
> D1<-ifelse(data$D==1, 1, 0)
> D2<-ifelse(data$D==-1, 1, 0)
> data1<-cbind(data[, -c(1, 4, 6)], D1, D2, GI)
> head(data1)
       V  I W      P  N D1 D2      GI
1 0.5168  1 0  4.252  3  1  0   2.229
2 0.3612  1 1 16.535  5  0  0 -11.463
3 0.4176 -1 0  5.161 10  0  1   3.872
4 0.4118 -1 0  0.183  7  0  0  -4.623
5 0.5916 -1 0  7.069  4  0  1  14.901
6 0.6246  1 0  2.362  9  1  0  11.921
> cor(data1[, -1])
            I          W           P          N         D1          D2          GI
I   1.0000000  0.3892495  0.11921266  0.2650860  0.7479576 -0.66332496  0.20560659
W   0.3892495  1.0000000  0.64831150  0.2718636  0.2401922 -0.25819889 -0.18685769
P   0.1192127  0.6483115  1.00000000 -0.1670507 -0.1036814  0.01942015 -0.38056872
N   0.2650860  0.2718636 -0.16705075  1.0000000  0.2824205 -0.20532004  0.29265096
D1  0.7479576  0.2401922 -0.10368142  0.2824205  1.0000000 -0.49613894  0.35454668
D2 -0.6633250 -0.2581989  0.01942015 -0.2053200 -0.4961389  1.00000000  0.02216247
GI  0.2056066 -0.1868577 -0.38056872  0.2926510  0.3545467  0.02216247  1.00000000

> sqrt(max(eigen(cor(data[,-1]))$values)/eigen(cor(data[,-1]))$values) #상태수
[1] 1.000000 1.095932 1.611830 1.993642 2.476476 3.752174 4.192147
```

```
> VIF(lm(V~I+D1+D2+W+GI+P+N, data=data1)) #VIF
       I       D1       D2        W       GI        P        N 
3.555492 2.678628 2.026857 2.724643 1.535663 2.612056 1.476388 
```

(b) (a)의 상관행렬, 상태지수, VIF를 확인하면, 강한 공선성의 증거를 찾기는 어려워 보인다. 하지만 변수 W와 P의 상관계수가 0.64이고, VIF가 각각 2.72, 2.61인 점을 통해 공선성의 존재를 의심할 수 있고, 변수 I 또한 $D1, D2$와 각각의 상관계수 0.74, -0.66를 갖고, 제일 큰 VIF인 3.55인 점을 통해 공선성이 발견될 수 있을 것으로 생각된다.
따라서 변수 W와 I를 제거하여, 결과를 확인해 보도록 하겠다.

```
> cor(data1[, -c(1, 2, 3)])
             P          N         D1          D2          GI
P   1.00000000 -0.1670507 -0.1036814  0.01942015 -0.38056872
N  -0.16705075  1.0000000  0.2824205 -0.20532004  0.29265096
D1 -0.10368142  0.2824205  1.0000000 -0.49613894  0.35454668
D2  0.01942015 -0.2053200 -0.4961389  1.00000000  0.02216247
GI -0.38056872  0.2926510  0.3545467  0.02216247  1.00000000

>sqrt(max(eigen(cor(data1[,-c(1, 2, 3 )]))$values)/eigen(cor(data1[,-c(1, 2, 3)]))$values)
#상태수
[1] 1.000000 1.248849 1.592816 1.688590 2.339708

> VIF(lm(V~D1+D2+GI+P+N, data=data1)) #VIF
      D1       D2       GI        P        N 
1.635715 1.441396 1.472593 1.179847 1.166061 
```

위의 상태지수에서 상태수가 2.33로 줄었음을 확인할 수 있다. 또한 변수 W와 공선성이 있을 것으로 생각되었던 P의 VIF가 크게 줄었고, 마찬가지로 변수 I와 공선성이 있을 것으로 생각된 D_1, D_2의 VIF 또한 크게 감소했음을 알 수 있다.
위의 상태수와 VIF의 감소를 통해, 다중공선성의 문제를 감소시켰다고 볼 수 있을 것이다.

(c) 위의 결과를 통해, 회귀모형이 $V = \beta_0 + \beta_1 P + \beta_2 N + \beta_3 (G \cdot I) + \alpha_1 D_1 + \alpha_2 D_2 + \epsilon$임을 알 수 있다. 회귀모형을 적합하는 과정은 다음과 같다.

```
> summary(lm(V~P+N+GI+D1+D2, data=data1))

Call:
lm(formula = V ~ P + N + GI + D1 + D2, data = data1)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.045233 -0.022414 -0.002339  0.016685  0.083647 
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.510337   0.029415  17.349 2.45e-11 ***
P           -0.001159   0.002725  -0.425 0.676516
N           -0.004904   0.003430  -1.430 0.173311
GI           0.008995   0.001881   4.783 0.000242 ***
D1           0.042919   0.023929   1.794 0.093060 .
D2          -0.029675   0.024147  -1.229 0.238023
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04164 on 15 degrees of freedom
Multiple R-squared:  0.7693,    Adjusted R-squared:  0.6923
F-statistic:    10 on 5 and 15 DF,  p-value: 0.0002314
```

따라서 적합된 회귀식은 $V = 0.510 - 0.001P - 0.004N + 0.008(G \cdot I) + 0.004D_1 - 0.03D_2 + \epsilon$이다.

(d) 9.4의 수정결정계수 R_a는 0.8248로 9.5의 0.6923보다 크다. 이는 9.4 회귀모형의 독립변수들이 종속변수 V를 9.5보다 잘 설명하고 있음을 나타낸다. 또한 잔차의 표준편차가 0.0314로 0.04164보다 근소하게 작아, 마찬가지로 9.4의 모형이 9.5의 모형보다 V를 잘 설명한다고 할 수 있을 것이다.