

## 4장 정리

표본 간차에서  $\mu$  가 2 or 3 보다 큰 값을 가지는 경우 특이값으로 결정

$$\text{지레점 } (P_{ii}) = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_k (x_k - \bar{x})^2} \text{ 이 } 2\mu = \frac{2(p+1)}{n} \text{ 보다 큰 값을 가지는 경우, 특이값으로 결정.}$$

$P_{ii} + \frac{e_i^2}{SSE} \leq 1$  이로  $P_{ii}$  가 높으면  $e_i^2$ 은 주어진다. 이는 승법모형에서 가변률체를 (2종오류) 모델에 써는 원인

$$Cook \text{의 거리} \quad r_i^2 = \left( \frac{e_i}{\hat{\sigma} \sqrt{P_{ii}}} \right)^2$$

$$C_i = \frac{r_i^2}{p+1} \times \frac{P_{ii}}{1-P_{ii}}, \text{ 기준 } F_{0.5}( (p+1), n-(p+1) ) \quad \text{보다 큰값일 때.}$$

$$DFITS_i = r_i^* \sqrt{\frac{n-p-2}{n-p-1-r_i^2}}$$

$$DFITS_i = r_i^* \sqrt{\frac{P_{ii}}{1-P_{ii}}}, \text{ 기준 } 2\sqrt{\frac{p+1}{n-(p+1)}} < |DFITS_i| \quad \text{보다 큰값.}$$

Hadi

$$H_i = \frac{P_{ii}}{1-P_{ii}} + \frac{p+1}{1-P_{ii}} \cdot \frac{d_i^2}{1-d_i^2}, \quad d_i^2 = \frac{e_i^2}{SSE}, \text{ 기준, 보다 큰값.}$$

결재성합수 P      간차합수 R



첨가 변수 풀로.

VS

성분 간차풀로.

회귀식이 쉽고

특정예측 변수 도입.

\* 3번트 회귀.  $\Sigma e_i^2$

특이값 찾기가 쉬움.

비선형성 예방.

$$\begin{cases} y_i - \beta_0 - \beta_1 x_i & , |y_i - \beta_0 - \beta_1 x_i| < c \\ \text{sign}(y_i - \beta_0 - \beta_1 x_i) \times C & , |y_i - \beta_0 - \beta_1 x_i| \geq c \end{cases}$$

예시)

$$x=13, y=12.74, x-\bar{x}=13-9=4.$$

$$P_{33} = \frac{1}{n} + \frac{(x_3 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{1}{11} + \frac{4^2}{110} = 0.24$$

$$C_3 = \frac{r_3^2}{p+1} \cdot \frac{P_{33}}{1-P_{33}} = \frac{3^2}{11} \cdot \frac{0.24}{1-0.24} = 1.42$$

$$\hat{y} = 3.002 + 0.5x. \Rightarrow \hat{y}_3 = 9.502$$

$$DFITS_3 = r_3^* \sqrt{\frac{P_{33}}{1-P_{33}}} = (1203.54) \sqrt{\frac{0.24}{1-0.24}}$$

$$= 676.3311$$

$$y - \hat{y}_3 = 12.74 - 9.502 = 3.238$$

$$r_3 = \frac{P_{33}}{\hat{\sigma} \sqrt{1-P_{33}}} = \frac{y - \hat{y}_3}{\hat{\sigma} \sqrt{1-P_{33}}} = \frac{3.238}{1.236 \sqrt{1-0.24}} = 2.999 \dots \approx 3$$

$$r_3^* = r_3 \sqrt{\frac{n-p-2}{n-(p+1)-r_3^2}} = 3 \sqrt{\frac{11-1-2}{11-(1+1)-9}} = 1203.54$$

$$\begin{aligned} H_3 &= \frac{P_{33}}{1-P_{33}} + \frac{1+p}{1-P_{33}} \cdot \frac{d_3^2}{1-d_3^2} \cdot \left( d_3 = \frac{e_3}{SSE} \right) \\ &= \frac{0.24}{1-0.24} + \frac{2}{1-0.24} \cdot \frac{(0.51)^2}{1-(0.81)^2} \end{aligned}$$

## 5장 정리

- 데이터 식별 후 질적변수를 가변수로 설정하여 나타내기

1. 가변수를 포함하지 않은식

2. 가변수를 포함한 가법모형식

3. 가변수를 포함한 승법모형식

$i (i=1,2,3)$  번 모형에 대해 표준화 전자그림여행 범주에 따른 관계그림을  
보는지에 짐작하지 않으면  $i+1$ 로 넘어가고, 짐작되면  $i$  번째 모형을  
적합식으로 결정한다.

~~가변수 or 대립종~~ : 가변수 값이 모두 0인 변수

- 두 집단의 비교.

$$\text{통합 } Y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij} \Rightarrow RM \quad k=1$$

$$\text{개별 소수 } Y_{ij1} = \beta_{01} + \beta_{11} x_{ij1} + \varepsilon_{ij1}$$

$$\text{백인 } Y_{ij2} = \beta_{02} + \beta_{12} x_{ij2} + \varepsilon_{ij2}$$

$$\Rightarrow \text{조건} \quad \beta_{01} = \beta_{02}, \quad \beta_{11} = \beta_{12}$$

(점화)

(기울기)

$$\text{가변수통합 } Y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma_i z_{ij} + \delta_i (x_{ij} \cdot z_{ij}) + \varepsilon_{ij} \Rightarrow FM \quad p=3 \text{ or } p=2$$

$$\text{가변수소수 } Y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma_i + \delta_i x_{ij} + \varepsilon_{ij}$$

$$= (\beta_0 + \gamma_i) + (\beta_1 + \delta_i) x_{ij} + \varepsilon_{ij}$$

$$\Rightarrow \text{조건} \quad \gamma_i = 0, \quad \delta_i = 0$$

(점화)

(기울기)

$$\text{가변수백인 } Y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}$$

~~대립집단 (모든 자시변수들이 0이다)~~

$$\therefore \text{가설 } H_0: (RM) \quad \dots k$$

$$Vs \quad H_1: (FM) \quad (r=0, s=0 \text{ 기준을 보고 계수정리}) \quad \dots p$$

$$F \text{ 결정통계량} = \frac{\text{SSE}(RM) - \text{SSE}(FM) / p-k}{\text{SSE}(FM) / n-(p+1)} = F \sim F_{0.05} (p-k, n-(p+1))$$

$$t \text{ 결정통계량} = \frac{\hat{\beta}}{se(\hat{\beta})} = t \sim t_{0.05} (n-(p+1)) \quad (\because p-k=1 \text{ 일 때 가능})$$

검정통계량 > 기각역 :  $H_0$  거짓.

\* 합격점수.

$$X_m = \frac{Y_m - \hat{\beta}_0}{\hat{\beta}_1} \quad \text{신뢰구간} \quad X_m \pm t_{\alpha/2} (n-2) \cdot \left( \frac{\hat{\sigma}}{\hat{\beta}_1} \right) \left( \frac{1}{n} \right)$$

- 계절성이 있는 경우.

계절성이 있는 시계열  $\rightarrow$  가법모형 가변수 모형  $\rightarrow$  승법모형 가변수모형.

순서대로 흐름에 맞게 진행.

$H_0: (RM)$  계절성 없는식.      Vs       $H_1: (IM)$  계절성 있는식.

## 9장 경리

예측변수간 강한 선형관계가 있는 경우 공선성이 있다고 흔히, 이를 가지고 분석하는 것은 의미가 없다.

### EEO 데이터.

$$\text{모형식. } ACHV = \beta_0 + \beta_1 \cdot FAM + \beta_2 \cdot PEER + \beta_3 \cdot SCHOOL + \varepsilon$$

- ① 진차 플랫. 인덱스 플랫을 보기  $\rightarrow$  톤이 값이 증가하는 것을 찾았지만 제거했을 때면 비슷한 대로 회귀값이 분석된다.
- ② 검정결과의  $R^2$ ,  $t$ 값,  $P$ 값 확인  $\Rightarrow R^2$ 은 작지만  $F$ 검정통계량으로 3개의 변수가 유의하다고 판별.
- ③ 공선성 확인  $\rightarrow$  예측변수들 사이의 상관계수가 크고, 개별  $t$ 검정값이 작은 걸로 보아 공선성의 증거가 된다.

### Import 데이터.

$$\text{모형식. } \text{Import} = \beta_0 + \beta_1 \cdot DOPROD + \beta_2 \cdot STOCK + \beta_3 \cdot CONSUM + \varepsilon$$

- ① 진차 플랫. 인덱스 플랫 확인  $\rightarrow$  추세를 보인다  $\rightarrow$  추세로  $R^2$ 은 크지만 개별  $t$ 검정값이 적으로 공선성을 확인  $\rightarrow$  데이터 조정.
  - ② 진차 플랫. 인덱스 플랫 확인  $\rightarrow$  절상적인 형태를 보인다.
  - ③ 검정결과의  $R^2$ ,  $t$ 값,  $P$ 값 확인  $\rightarrow R^2$ 이 크게 나왔지만 DOPROD의  $t$ 검정 값이 적으로 공선성 확인  $\rightarrow$  상관계수 확인(1) CONSUM과  $r = 0.97$  이었으며  $CONSUM = \frac{2}{3} DOPROD$ 였다.  $\rightarrow$  이 관계가 유지된다면  $R^2$ 이 크므로 예측하는데 지침은 없지만, 변수를 고정시킬 때 상관관계 있는 변수는 둘째에서 처리하자.
- $$\text{IMPORT}_{(1960)} = \text{IMPORT}_{(1959)} - 0.051 \times (10) \Rightarrow -0.51 \Rightarrow \text{나쁜 추정}$$
- $$\text{IMPORT}_{(1960)} = \text{IMPORT}_{(1959)} - 0.051 \times (10) + 0.281 \times (10 \times \frac{2}{3}) \Rightarrow 1.5 \Rightarrow \text{좋은 추정}$$

### adver 데이터.

$$\text{모형식 } S_t = \beta_0 + \beta_1 A_t + \beta_2 P_t + \beta_3 E_t + \beta_4 A_{t-1} + \beta_5 P_{t-1} + \varepsilon$$

- ① 진차 플랫. 인덱스 플랫을 보기  $\rightarrow$  절상적이다.
- ② 검정결과의  $R^2$ ,  $t$ 값,  $P$ 값, 상관계수 확인  $\rightarrow$  개별 상관계수가 다 크고  $R^2 = 0.97$ 로 높다. 그래서  $A_{t-1}, P_{t-1}$ 의  $t$ 검정값이 적음  $\rightarrow$  2개의 상관관계가 아닌 4개의 ( $A_t, P_t, A_{t-1}, P_{t-1}$ ) 관계가 존재했으며  $A_{t-1}, P_t, P_{t-1}$ 은  $A_t$ 를 97% 설명이 가능함을 확인할 수 있다.
- ③ 공선성이 확인되었다.

\* 진차그림. 인덱스 표준화진차 그림. 꺽임상관관계 그림은 비정확하다. 따라서 VIF(분산적대인자), 상관계수, 상수의 합 등을 활용해 공선성을 확인한다.

~~★~~

① 불산학적 인자 (VIF),  $D^2$ ,  $\overline{VIF}$

$$VIF = \frac{1}{1 - R_j^2}, (R_j^2 : lm(x_1 \sim x_1 + x_2 + \dots + x_{j-1}) \text{의 결정계수})$$

$$0(\text{비상관}) < R_j^2 < 1(\text{상관}) \Rightarrow \text{기준 } 0.9$$

$$1(\text{비상관}) < VIF_j < \infty(\text{상관}) \Rightarrow \text{기준 } 10.$$

$D^2$ : OLS 추정량과 그 축소값으로부터의 제곱거리, 주요수록 좋은 추정량이다.

$\overline{VIF}$ : 데이터가 치환하는 때 OLS 추정량의 상대적인 오차제곱의 크기, 주요수록 좋은 추정량이다.

$$D^2 = \sum_{j=1}^P VIF_j$$

$$\overline{VIF} = \frac{D^2}{P\tau^2} \quad (D^2 \text{가 모두 치환하는 때 }) = \frac{\sum VIF_j}{P\tau^2} = \frac{\sum VIF_j}{P} \quad ex) \overline{VIF} = 2 \text{ 일 때 해석: OLS 추정량의 저작오차가 예측변수들이 모두 치환할 때보다 } 2 \text{ 배 크다고 할 수 있다.}$$

② 상대지수, 상대수, 역수의 합, 고윳값

변수들 간의 상관계수 행렬을 만들고 그 행렬로 고윳값을 계산한다. (크기별로 나열)  $\Rightarrow \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ .

$\lambda_p \approx 0$  이면 공선성 존재  $\Rightarrow$  기준 0.05~0.1 (다음)

$\sum_{i=1}^p \frac{1}{\lambda_i}$  이 크면 공선성 존재  $\Rightarrow$  기준  $P \times 5$

$\sqrt{\frac{\lambda_1}{\lambda_p}}$  가 크면 공선성 존재  $\Rightarrow$  기준 15 (30)

$\sqrt{\frac{\lambda_1}{\lambda_p}} = k_1$  가 크면 공선성 존재  $\Rightarrow$  기준 15 (30)  $\Rightarrow$  여러 개의 집합이 생길 수 있음.

③ 고유벡터

데이터의 공선성 유무는 주된 핵심으니 어떤 변수들이 상관성을 가지고 있는지 확인해본다.

각 고윳값에 해당하는 고유벡터를 뿐으면 다음과 같다.

$V = (V_1, V_2, V_3, \dots, V_p) = \begin{bmatrix} V_{11} & V_{12} & \dots & V_{1p} \\ V_{21} & V_{22} & \dots & V_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ V_{p1} & V_{p2} & \dots & V_{pp} \end{bmatrix} \Rightarrow \lambda_1 \text{ 가 공선성이 있는 그룹이라면 다음과 같다.}$

$$\lambda_1 \approx V_{11} \tilde{x}_1 + V_{12} \tilde{x}_2 + \dots + V_{1p} \tilde{x}_p$$

그리고 0에 가까운 값들을 0으로 수렴시키면 상관성을 가지는 변수들이 나온다.

### EEO 데이터

$$\begin{array}{lll} \lambda_1 = 2.952 & \lambda_2 = 0.04 & \lambda_3 = 0.008 \\ k_1 = 1 & k_2 = 8.59 & k_3 = 19.26 \end{array} \quad \left. \begin{array}{l} \lambda_3 = V_{13}\tilde{X}_1 + V_{23}\tilde{X}_2 + V_{33}\tilde{X}_3 \\ 0 = V_{13}\tilde{X}_1 + V_{23}\tilde{X}_2 + V_{33}\tilde{X}_3 \end{array} \right\} \Rightarrow 3개의 변수가 상관성을 가짐. \\ \therefore k_3 > 15 이므로 공선성 존재. \end{math>$$

### IMPORT 데이터

$$\begin{array}{lll} \lambda_1 = 1.999 & \lambda_2 = 0.998 & \lambda_3 = 0.003 \\ k_1 = 1 & k_2 = 1.42 & k_3 = 21.26 \end{array}$$

$$k_3 > 15 이므로 공선성 존재 \rightarrow \lambda_3 = V_{13}\tilde{X}_1 + V_{23}\tilde{X}_2 + V_{33}\tilde{X}_3$$

$$V = \begin{bmatrix} 0.706 & -0.036 & -0.701 \\ 0.044 & 0.999 & -0.001 \\ 0.701 & -0.026 & 0.701 \end{bmatrix} \quad \begin{array}{l} 0.036 \stackrel{\circ}{=} -0.701\tilde{X}_1 - 0.001\tilde{X}_2 + 0.701\tilde{X}_3 \\ 0 \stackrel{\circ}{=} -0.701\tilde{X}_1 + 0.701\tilde{X}_3 \end{array}$$

$\tilde{X}_1 \stackrel{\circ}{=} \tilde{X}_3 \Rightarrow 2개(X_1, X_2)의 변수가 연관성을 가짐.$

### Adver 데이터

$$\begin{array}{lllll} \lambda_1 = 1.101 & \lambda_2 = 1.288 & \lambda_3 = 1.145 & \lambda_4 = 0.859 & \lambda_5 = 0.007 \\ k_1 = 1 & k_2 = 1.14 & k_3 = 1.21 & k_4 = 1.40 & k_5 = 15.29 \end{array}$$

$$k_5 > 15 이므로 공선성 존재$$

$$V = \begin{bmatrix} & -0.514 & & & \\ & -0.489 & & & \\ V_1 & V_2 & V_3 & V_4 & \begin{array}{l} 0.007 \stackrel{\circ}{=} -0.514\tilde{X}_1 - 0.489\tilde{X}_2 + 0.010\tilde{X}_3 - 0.428\tilde{X}_4 - 0.559\tilde{X}_5 \\ 0 \stackrel{\circ}{=} -0.514\tilde{X}_1 - 0.489\tilde{X}_2 - 0.428\tilde{X}_4 - 0.559\tilde{X}_5 \end{array} \\ 0.010 \\ -0.428 \\ -0.559 \end{bmatrix} \Rightarrow 4개(X_1, X_2, X_3, X_4)의 변수가 연관성을 가짐.$$

## 10장 정리

예측 변수들 사이에 공선성이 있는 경우 통상적인 최소제곱추정치는 불안정하고, 예측 변수를 무작위 제거하는 것 또한 잘못된 추론을 가져온다. 따라서 ① 고려 변수에 제약을 두거나, ② 최소제곱법 대안으로 ③ 주성분 고리와 ④ 능형회기 방정식을 이용한다.

\* 주성분:  $P$  개의 변수들의 집합을 척도화하는  $P$  개의 서로운 축들로 나타낸다.

$$V = [V_1 \ V_2 \ \dots \ V_p], \ C_j = V_{1j} \tilde{x}_1 + V_{2j} \tilde{x}_2 + \dots + V_{pj} \tilde{x}_p, \ j=1, 2, \dots, p$$

$$= \begin{bmatrix} V_{11} & V_{12} & \dots & V_{1p} \\ V_{21} & V_{22} & \dots & V_{2p} \\ \vdots & \ddots & \ddots & \vdots \\ V_{p1} & V_{p2} & \dots & V_{pp} \end{bmatrix} \quad \text{Var}(C) = \begin{bmatrix} \lambda_1 & & & \\ & \ddots & & \\ & & \lambda_p & \\ & & & 0 \end{bmatrix}$$

이 성립한다.

(a) Import 데이터

$$\Gamma(\text{삼관행렬}) = \begin{bmatrix} 1 & 0.026 & 0.991 \\ 0.026 & 1 & 0.036 \\ 0.991 & 0.036 & 1 \end{bmatrix}$$

$$\text{주성분 } C_1 = 0.106 \tilde{x}_1 + 0.044 \tilde{x}_2 + 0.701 \tilde{x}_3$$

$$C_2 = -0.036 \tilde{x}_1 + 0.999 \tilde{x}_2 - 0.26 \tilde{x}_3$$

$$C_3 = -0.701 \tilde{x}_1 - 0.001 \tilde{x}_2 + 0.101 \tilde{x}_3$$

$$\lambda_i(\text{고유값}) = [1.999 \ 0.998 \ 0.003], \ \sum \lambda_i = P$$

$$= V^T \cdot Data_S$$

$$V(\text{고유벡터}) = \begin{bmatrix} 0.106 & -0.036 & -0.701 \\ 0.044 & 0.999 & -0.001 \\ -0.701 & -0.26 & -0.001 \end{bmatrix}$$

$$(\text{주성분 }) C = \begin{bmatrix} -2.126 & 0.639 & -0.021 \\ -1.619 & 0.556 & -0.011 \\ \vdots & \vdots & \vdots \\ 1.931 & -1.663 & -0.081 \end{bmatrix}$$

$$Data_S(\text{표본데이터}) = \begin{bmatrix} -1.5091 & 0.5457 & -1.5332 \\ -1.1130 & 0.4851 & -1.2085 \\ \vdots & \vdots & \vdots \\ 1.4803 & -1.5765 & 1.3503 \end{bmatrix}^T$$

$$(\text{주성분들의 } \text{방수공분산행렬}) Cov = \begin{bmatrix} 1.999 & 0 & 0 \\ 0 & 0.998 & 0 \\ 0 & 0 & 0.003 \end{bmatrix}$$

$\tilde{x}_1 \ \tilde{x}_2 \ \tilde{x}_3$

$\lambda_i$ 가 0이 가까워지면  $C_i$ 의 분산이 0이 가까워진다는 뜻이고 이는 변수가 정교진다는 뜻이므로 상관성이 끊어진다.

$$\text{Var}(C_3) = \lambda_3 = 0.003 \approx 0 \text{ 이므로 이는 곧 } C_3 \text{의 평균값이 될 것이다.}$$

$$\Rightarrow C_3 = -0.701 \tilde{x}_1 - 0.001 \tilde{x}_2 + 0.101 \tilde{x}_3 \approx 0 \Rightarrow \tilde{x}_1 \approx \tilde{x}_3 \text{ (이는 두 변수 상관성이 } R = 0.991 \text{이 부합하는 값)}$$

04) adver 예제

$$r(\text{성분계수}) = \begin{bmatrix} 1 & -0.357 & -0.129 & -0.14 & -0.496 \\ -0.357 & 1 & 0.063 & -0.316 & -0.296 \\ -0.129 & 0.063 & 1 & -0.166 & 0.208 \\ -0.14 & -0.316 & -0.166 & 1 & -0.358 \\ -0.496 & -0.296 & 0.208 & -0.358 & 1 \end{bmatrix}$$

$$\hat{\gamma}_5(\text{고유값}) = \begin{bmatrix} 1.101 & 1.288 & 1.145 & 0.859 & 0.001 \end{bmatrix}$$

$$V(\text{고유분수}) = \begin{bmatrix} 0.532 & 0.024 & 0.668 & -0.074 & -0.514 \\ -0.232 & -0.825 & -0.158 & 0.037 & -0.489 \\ -0.389 & 0.022 & 0.217 & -0.895 & 0.010 \\ 0.395 & 0.260 & -0.692 & -0.338 & -0.428 \\ -0.596 & 0.501 & 0.057 & 0.219 & -0.559 \end{bmatrix}$$

$$\text{주성분식 } C_1 = 0.532\tilde{x}_1 - 0.232\tilde{x}_2 - 0.389\tilde{x}_3 + \dots - 0.596\tilde{x}_5$$

$$C_5 = -0.514\tilde{x}_1 - 0.489\tilde{x}_2 + 0.010\tilde{x}_3 + \dots - 0.559\tilde{x}_5$$

(주성분들의  
분산공분산행렬)

$$\text{Cor} = \begin{bmatrix} 1.101 & 0 & 0 & 0 & 0 \\ 0 & 1.288 & 0 & 0 & 0 \\ 0 & 0 & 1.145 & 0 & 0 \\ 0 & 0 & 0 & 0.859 & 0 \\ 0 & 0 & 0 & 0 & 0.001 \end{bmatrix}$$

$\therefore$  즉 최소고유값  $\hat{\gamma}_5 = 0.001 \approx 0$  이므로  $C_5$ 의 평균은 0으로 성립과 푼다.

$$\tilde{x}_5^* \doteq -0.514\tilde{x}_1 - 0.489\tilde{x}_2 + 0.010\tilde{x}_3 - 0.428\tilde{x}_4 - 0.559\tilde{x}_5$$

$$0 \doteq -0.514\tilde{x}_1 - 0.489\tilde{x}_2 - 0.428\tilde{x}_4 - 0.559\tilde{x}_5 \Rightarrow x_1 \doteq -0.951\tilde{x}_2 - 0.833\tilde{x}_4 - 1.081\tilde{x}_5$$

$\therefore$  따라서  $x_1, x_2, x_4, x_5$  가 연관성을 가지고 있다.

$$\text{모형식 } y_{ij} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$$

$$\text{표현화변수식 } \tilde{y} = \theta_1 \tilde{x}_1 + \theta_2 \tilde{x}_2 + \theta_3 \tilde{x}_3 + \theta_4 \tilde{x}_4 + \theta_5 \tilde{x}_5 + \varepsilon'$$

$$\text{주성분사용 } \tilde{y} = \alpha_1 C_1 + \alpha_2 C_2 + \alpha_3 C_3 + \alpha_4 C_4 + \alpha_5 C_5 + \varepsilon'$$

$$\text{교착지. (2번수준 예시)} \\ \tilde{y} = \theta_1 \tilde{x}_1 + \theta_2 \tilde{x}_2$$

$$\therefore \hat{\theta}_i = \frac{s_y}{s_i} \theta_i \quad \hat{\beta}_i = \bar{y} - \hat{\theta}_1 \bar{x}_1 - \hat{\theta}_2 \bar{x}_2$$

$$\frac{y-\bar{y}}{s_y} = \theta_1 \frac{x_1 - \bar{x}_1}{s_1} + \theta_2 \frac{x_2 - \bar{x}_2}{s_2}$$

$$y - \bar{y} = \frac{s_y}{s_1} \theta_1 x_1 + \frac{s_y}{s_2} \theta_2 x_2 - \frac{s_y}{s_1} \theta_1 \bar{x}_1 - \frac{s_y}{s_2} \theta_2 \bar{x}_2$$

$$y = (\bar{y} - \frac{s_y}{s_1} \theta_1 \bar{x}_1 - \frac{s_y}{s_2} \theta_2 \bar{x}_2) + \frac{s_y}{s_1} \theta_1 x_1 + \frac{s_y}{s_2} \theta_2 x_2 = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

① 주성분계산

$$C = \begin{bmatrix} C_1 \\ \vdots \\ C_5 \end{bmatrix} = \begin{bmatrix} V_1 \\ \vdots \\ V_5 \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_5 \end{bmatrix} \quad 5 \times 1 \quad 5 \times 5 \quad 5 \times \text{now (data\_s)}$$

$$C = V^T \cdot \tilde{X} \quad \tilde{X} = V C \quad \xrightarrow{\text{기호}}$$

-  $\hat{\theta}$  와  $\hat{\alpha}$  의 관계식.

$$\hat{\theta} = \begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \vdots \\ \hat{\theta}_5 \end{bmatrix} = \begin{bmatrix} V_{11} & V_{12} & \dots & V_{15} \\ V_{21} & V_{22} & \dots & V_{25} \\ \vdots & \vdots & \ddots & \vdots \\ V_{51} & V_{52} & \dots & V_{55} \end{bmatrix} \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \vdots \\ \hat{x}_5 \end{bmatrix}, \quad \hat{\theta}_4 = 0.395 \\ = V_{41} \hat{x}_1 + \dots + V_{45} \hat{x}_5 \\ = 0.394912$$

$$\therefore \hat{\theta} = V^T \hat{\alpha} \quad \text{설명.}$$

② 추정회귀 계산.

기준

변수 계수 표준오차

$$C_1 -0.3653 0.053 R^2=0.97$$

$$C_2 -0.4169 0.064 \hat{T}=0.3303$$

$$C_3 0.1619 0.067 n=22$$

$$C_4 -0.7036 0.078 df=n-(p+1) \\ = 16$$

$$C_5 -1.2192 0.0846$$

변수 계수 표준오차

$$\hat{x}_1 0.583 0.438 R^2=0.97$$

$$\hat{x}_2 0.913 0.417 \hat{T}=0.3303$$

$$\hat{x}_3 0.786 0.075 n=22$$

$$\hat{x}_4 0.395 0.361 df=16$$

$$\hat{x}_5 0.503 0.416$$

-  $\hat{\alpha}$  구하는 방법.

표현화된  $Y$ 에 주성분의 예측변수들을 적용  
시킨 시기변수에서 추정량.

$$\hat{\alpha} = [-0.3653 \dots -1.2192]$$

-  $\hat{\theta}$  구하는 방법

표현화된  $Y$ 에 표현화된  $X$ 들을 적용시킨 기계방식  
이 추정법

$$\hat{\theta} = [0.583 \dots 0.503]$$

## 포증오차 구하는 방법.

$$\begin{bmatrix} \text{s.e.}(\hat{\theta}_1) \\ \vdots \\ \text{s.e.}(\hat{\theta}_k) \end{bmatrix} = \begin{bmatrix} V \\ \vdots \\ V \end{bmatrix} \begin{bmatrix} \text{s.e.}(\hat{\theta}_1) \\ \vdots \\ \text{s.e.}(\hat{\theta}_k) \end{bmatrix}^T \begin{bmatrix} V \\ \vdots \\ V \end{bmatrix} \begin{bmatrix} \text{s.e.}(\hat{\theta}_1) \\ \vdots \\ \text{s.e.}(\hat{\theta}_k) \end{bmatrix}$$

제막의 부사

예) import E101E1

$$\text{Import} = \beta_0 + \beta_1 \text{DOPROD} + \beta_2 \text{STOCK} + \beta_3 \text{CONSUM} + \varepsilon_i$$

$$= \beta_0 + \beta_1 (\text{DORRDP} + \text{CONSUM}) + \beta_2 \text{STOCK} + \varepsilon_i \Rightarrow 24105 \quad (\beta_1 = \beta_2)$$

$$= \beta_0 + \beta_2 STOCK + \beta_1 NEWVAR + \varepsilon_i$$

즉 차원less 그림에서 이상이 없고 회귀분석에서  $R^2 = 0.987$  이 성공계수도 적으며 두 변수의 t검정 값도 높아 잘 적합되었다고 할 수 있다.

$$IMPORT = -9.007 + 0.0086 \cdot DOPROD + 0.612 \cdot STOCK + 0.086 \cdot CONSUM + e.$$

$$H_0: (\text{RM}) \quad \text{IMPORT} = \beta_0 + \beta_2 \text{STOCK} + \beta_1 \text{NEWVAR} + \varepsilon; \quad k=2$$

$$H_1: (FM) \text{ IMPORT} = \beta_0 + \beta_1 DOPROD + \beta_2 STOCK + \beta_3 CONSUM + \varepsilon_i, \quad p=3$$

$$F_{\text{검정 통계량}} = \frac{(2.543 - 1.613)/1}{1.613 / (1 - (3+1))} = 3.8493, \quad F_{\alpha}(1, 7) = 5.0215$$

$\therefore H_0$  제적  $\Rightarrow$  저영구진  $\beta_1 = \beta_3$ 은 유의한다고 볼수 있다.

$$\text{Import} = \beta_0 + \beta_1 \text{DOPROD} + \beta_2 \text{STOCK} + \beta_3 \text{CONSUM} + \epsilon_i$$

$$= \beta_0 + (\beta_1 + \frac{2}{3}\beta_3) DOPROD + \beta_2 STOCK + \varepsilon_i \leftarrow Z(19). \quad (Consum = \alpha_1 + \frac{2}{3}DOPROD)$$

위의 두 종류의 주 저항이 250을 넘을 수 있다. 또한 DOPROD 와 STOCK 의 상관계수는 0.026으로 긴밀성이 있는 것으로 확인되었다.

DOPROD의 계수는 수수료 DOPROD의 영향이 아니며 consumo의 일부로 고려된다.

\* 광진성이 있는 데이터 처리

다중금전체성이 있는 경우 혁결책으로 최소제곱법 대신 주성분 회귀와 능형회귀방식을 이용하는 것으로 제안된다.

즉성별 인지 ————— 예를 병수들의 비직교성으로 허락

【】 키워드들이 블여지는 제약조직에 관련한 해석

능형 지지 — 기관계수들에 부여되는 제약조건에 관련한 허가

~~두 가지의 추정량은~~ 편향이 있고, OLS 보다 SSE 가 크고  $R^2$  도 낮다.

하지만  $\text{SSE}$ (평균제곱오차) 관점에서  $OLS$ 를 찾았다 높은 점집도를 가지고 있다.

## \* 주성분 고지

예) Import 데이터

$$\text{IMPORT} = \beta_0 + \beta_1 \text{DOPROD} + \beta_2 \text{STOCK} + \beta_3 \text{CONSUM} + \varepsilon_i$$

$$\tilde{Y} = \alpha_1 \tilde{X}_1 + \alpha_2 \tilde{X}_2 + \alpha_3 \tilde{X}_3 + \varepsilon'$$

표준화된  $\tilde{Y}$ 에 표준화된 예측변수를 적합한 회귀분석

계수	S.E	t정점	P값	$R^2 = 0.992$
$\tilde{X}_1$	-0.339	0.464	.	$\hat{\alpha} = 0.034$
$\tilde{X}_2$	0.213	0.034	.	$df = n - (p+1) = 7$
$\tilde{X}_3$	1.303	0.464	.	$n = 11$

표준화된  $\tilde{Y}$ 에 주성분 예측변수를 적합한 회귀분석

계수	S.E	t정점	P값	$R^2 = 0.992$
$C_1$	0.696	0.024	.	$\hat{\alpha} = 0.034$
$C_2$	0.191	0.034	.	$df = 7$
$C_3$	1.160	0.656	.	$n = 11$

$$\hat{\theta} = \begin{bmatrix} -0.339 \\ 0.213 \\ 1.303 \end{bmatrix} = \begin{bmatrix} 0.706 & -0.036 & -0.701 \\ 0.044 & 0.999 & -0.007 \\ 0.701 & -0.026 & 0.701 \end{bmatrix} \begin{bmatrix} 0.696 \\ 0.191 \\ 1.160 \end{bmatrix} = V \hat{\alpha}$$

$$\begin{aligned} \tilde{Y} &= \alpha_1 C_1 + \alpha_2 C_2 + \alpha_3 C_3 + \varepsilon' \\ &= 0.690 C_1 + 0.191 C_2 + 1.160 C_3 + \varepsilon' \end{aligned}$$

⇒ 설명력이 떨어지는 주성분은 제거하여 차원을 조정함.

$$\text{주성분식. } C_1 = 0.706 \tilde{X}_1 + 0.044 \tilde{X}_2 + 0.701 \tilde{X}_3$$

$$C_2 = -0.036 \tilde{X}_1 + 0.999 \tilde{X}_2 - 0.026 \tilde{X}_3$$

$$C_3 = -0.701 \tilde{X}_1 - 0.007 \tilde{X}_2 + 0.701 \tilde{X}_3$$

$$\lambda_1 = 1.999 \quad \lambda_2 = 0.999 \quad \lambda_3 = 0.002 \Rightarrow \lambda_3 \approx 0 \text{ 이므로 } C_3 \text{ 제거.}$$

$$\tilde{Y} = \alpha_1 C_1 + \alpha_2 C_2 + \alpha_3 C_3 + \varepsilon' \quad (M1)$$

$$\tilde{Y} = \alpha_1 C_1 + \alpha'_2 C_2 + \varepsilon' \quad (M2)$$

$$\tilde{Y} = \alpha''_1 C_1 + \varepsilon' \quad (M3)$$

$C_3$ 가 제거되었으므로 추천되는 모형은 M2, M3이다.  
단, M2, M3는 주성분을 제거한 모형으로 편학이 있을 수 있다.  
또한  $C_1, C_2, C_3$ 는 모두 정규분포로  $\alpha_1 = \alpha'_1 = \alpha''_1$ ,  $\alpha_2 = \alpha'_2$  가 성립한다.  
하지만  $\varepsilon'$ 에 대해서는 표준화변수는 적합하지 않으므로  $\hat{\theta}$ 는 주의해야 한다.

$$\hat{\theta} = \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\alpha}_3 \end{bmatrix} = \begin{bmatrix} V \end{bmatrix} \begin{bmatrix} \hat{\alpha} \end{bmatrix} \quad (\text{단 } M1 : \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\alpha}_3 \end{bmatrix}, M2 : \begin{bmatrix} \hat{\alpha}'_1 \\ \hat{\alpha}'_2 \\ 0 \end{bmatrix}, M3 : \begin{bmatrix} \hat{\alpha}''_1 \\ 0 \\ 0 \end{bmatrix}) \quad \hat{\alpha} = \begin{bmatrix} 0.69 \\ 0.191 \\ 1.154 \end{bmatrix} \text{ 을 통해 적합화되었던.}$$

$$M_3 \quad \tilde{Y} = -0.339 \tilde{X}_1 + 0.213 \tilde{X}_2 + 1.302 \tilde{X}_3$$

$$M_2 \quad \tilde{Y} = 0.4805 \tilde{X}_1 + 0.2211 \tilde{X}_2 + 0.4805 \tilde{X}_3$$

$$M_1 \quad \tilde{Y} = 0.4814 \tilde{X}_1 + 0.03 \tilde{X}_2 + 0.4815 \tilde{X}_3$$

$X_1$ 의 계수가 징역입니다

$X_2$ 의 계수가 이상입니다.

\* 즉 M3는 공산성을 포함하고 있고

M1은  $X_2$ 의 설명력이 부족으로

M2를 적합한다.

DOPROD와 CONSUM 사이에는  $DOPROD = \alpha_1 + 0.69 \text{ CONSUM}$ 의 관계가 일치함으로 DOPROD의 주변인자로 CONSUM의 주변인자의 69%가 짐을 의미한다.

즉 성분 고지에는 대중공선성 제거에 효과적이지만 B를 데이터에서 효과적인 것은 아니다. 또한 주성분 고지는 높은 지지점 ( $R_{\text{adj}}$ )을 가지는 등이 있어 대해 상당히 영향을 끌 수 있다.

## \* 능형회귀와 가쏘회귀

능형회귀는 비직교적일 때 사용하는 방법이다.

다중선형회귀에서  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$  는 최소제곱법을 통해 회귀계수들의 최소값을 찾았다.

$$S(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p) = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \dots - \hat{\beta}_p x_p)^2$$

능형회귀는 OLS 와 두의 합을 볼에서 최소화한다는 점에서 차이가 있다. (조율)

$$\sum (y_i - \beta_0 - \beta_1 x_1 - \dots - \beta_p x_p)^2 + \lambda \sum_{i=1}^p \beta_i^2$$

즉 소별점 항이라고 하며, OLS 보다 추정량을 0으로 가깝게 놓여주기 때문이다.  
 $\lambda = 0$  일 때 OLS 와 동일,  $\lambda = \infty$  일 때  $\beta_1 = \beta_2 = \dots = \beta_p = 0$  이 될 것이다.

가쏘는 능형회귀의 별점과 향으로 인하여 원본하는 0을 뺏는들이 내는 단점을 보완하기 위해 도입.

$$\sum (y_i - \beta_0 - \beta_1 x_1 - \dots - \beta_p x_p)^2 + \lambda |\beta_1|$$

능형회귀와 다르게 흔해서 만날 가능성을 높여주므로 0의 값을 만드는 가능성을 높인다.

그럼까지 배운 OLS로 얻은 추정량은 불편추정량으로, 불편추정량 중 분산이 가장 작은 성질을 가지고 있다 (BLUE)

하지만 능형회귀와 가쏘는 불편추정량이 아니므로 양간의 편향성이 생기지만 OLS보다 더 작은 분산을 가진다.

## 11장 정리

$$(FM) \quad \hat{Y}_{ij} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \dots + \beta_q x_{iq} + \varepsilon \quad \dots (M1)$$

$$(RM) \quad \hat{Y}_{ij} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon \quad \dots (M2)$$

$$\hat{Y}_i^* = \hat{\beta}_0^* + \hat{\beta}_1^* x_{i1} + \hat{\beta}_2^* x_{i2} + \dots + \hat{\beta}_q^* x_{iq}$$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip}$$

라고 하면 때, (FM) 아닌  $\hat{\beta}_j^*$ 이 불편추정량이지만  $\hat{\beta}_j$ 는 편향추정량이다.  $\Rightarrow E(\hat{\beta}_j^*) = \beta_j \neq E(\hat{\beta}_j)$

대신  $\hat{\beta}_j$ 의 분산은  $\hat{\beta}_j^*$ 보다 작다.

$$Var(\hat{\beta}_j^*) \geq Var(\hat{\beta}_j)$$

$$Var(\hat{Y}_i^*) \geq Var(\hat{Y}_i)$$

$$\begin{aligned} MSE(\hat{\beta}_j^*) &= E[(\hat{\beta}_j^* - \beta_j)^2] = E[(\hat{\beta}_j^* - E(\hat{\beta}_j^*) + E(\hat{\beta}_j^*) - \beta_j)^2] \\ &= E[(\hat{\beta}_j^* - E(\hat{\beta}_j^*))^2] + E[(E(\hat{\beta}_j^*) - \beta_j)^2] = E[(\hat{\beta}_j^* - E(\hat{\beta}_j^*))^2] + [E(\hat{\beta}_j^*) - \beta_j]^2 \\ &= Var(\hat{\beta}_j^*) + [Bias(\hat{\beta}_j^*)]^2 \end{aligned}$$

$\therefore$  FM: 불편추정량, BLUE.

$\therefore$  RM: 편향추정량. OLS 보다 분산이 작음.

유일하게 "가장 좋은" 예측변수들의 집단은 있을 수 있다. 여러개의 가능한 변수집단을 고르는 것이 바람직하다.  
회귀방법식을 평가하기 위한 방법들이다.

### • 진짜평균 제곱 (RMS)

$$RMS_p = \frac{SSE_p}{n-p} \quad (SSE_p: 진짜제곱합) \rightarrow RMS_p 가 진짜것을 추적한다.$$

유의미한 변수가 추가되면  $SSE_p$ 의 값은 커져  $RMS_p$ 가 감소하게 되고, 무의미한 변수가 추가되면 오히려  $RMS_p$ 가 증가하게 된다.

$$R_p^2 = 1 - (n-p) \frac{RMS}{SST}$$

$$R_{ap}^2 = 1 - (n-1) \frac{RMS}{SST}$$

### • Mallows C<sub>p</sub>

전체의 예측변수 중 일부를 사용하면 예측값은 일반적으로 편향되어 있으며, 따라서 0이 사용해 평가한다면 분산보다 MSE를 고려해야 한다.

$$J_p = \frac{1}{\tau^2} \sum_{i=1}^n MSE(\hat{Y}_{ij}) \quad (\tau^2: 편차합의 분산)$$

$$\hat{J}_p = C_p \quad (= J_p의 추정값) = \frac{SSE_p}{\tau^2} + (2p-n), \quad (p는 선형변수)$$

$C_p$ 는  $p$ 에 가까울수록 비강제한 예측변수들의 집합을 나타냄.

### • AIC / BIC / AIC<sup>c</sup>

$$AIC_p = n \log \left( \frac{SSE_p}{n} \right) + 2p \quad (p는 상수함 포함 계수의 수)$$

$$BIC_p = n \log \left( \frac{SSE_p}{n} \right) + p \log n \quad * AIC의 큰 p를 결정하려는 경향을 제거. n > 8 일때부터 차이가 낸다.$$

두 추정과 결론값이 없어야 하며, 값이 2정도 차이가 낸다면 통일적으로 높다.

$$AIC_p^c = AIC_p + \frac{2(p+2)(p+3)}{(n-p-3)} \quad \therefore 평의를 수정한 AIC$$

$\Rightarrow$  다중공선성을 확인하기 위해 VIF, 상수수,  $\frac{1}{R^2}$ ,  $R^2$ 를 확인한다.

예측변수가 많다고 하면 2<sup>nd</sup> 차의 방정식  $\rightarrow$  3, 4차의 좋은 방정식을 추려내기 위해  $R^2, C_p, RMS, AIC, BIC$  등을 사용할 수 있으며, FS, BE, 단계적 방법 등으로 변수선택을 흐름적으로 진행해본다.

### • FS (점진적 선택)

상수함만 있는 모형에서 출발.

$$|t| < t_{0.05}(n-p) \quad \text{즉은값들을 추가한다. } t = \frac{\hat{y}_i}{se(\hat{y}_i)}$$

### • BE

full 모형에서 출발

공현도가 작은 값부터 삭제하다가 모든 변수가 유의할 때 중단.  
(|t| < t<sub>0.05</sub>)

변수	min t	RMS	C <sub>p</sub>	rank	AIC	BIC	변수	min t	RMS	C <sub>p</sub>	rank		
X <sub>1</sub>	9.74	6.993	1.41	1	118.63	121.43	X <sub>1</sub> X <sub>3</sub> X <sub>6</sub> X <sub>2</sub> X <sub>4</sub> X <sub>5</sub>	0.26	7.068	7.07	1	123.36	133.17
X <sub>1</sub> X <sub>3</sub>	1.51	6.811	1.11	1	118.00	122.21	X <sub>1</sub> X <sub>3</sub> X <sub>6</sub> X <sub>2</sub> X <sub>4</sub>	0.49	6.928	5.01	1	121.43	129.86
X <sub>1</sub> X <sub>3</sub> X <sub>6</sub>	1.29	6.134	1.60	1	118.14	123.14	X <sub>1</sub> X <sub>3</sub> X <sub>6</sub> X <sub>2</sub>	0.59	6.820	3.28	1	119.73	126.73
X <sub>1</sub> (X <sub>3</sub> X <sub>6</sub> X <sub>2</sub> )	0.59	6.820	3.28	1	119.13	126.73	X <sub>1</sub> X <sub>3</sub> X <sub>6</sub>	1.29	6.734	1.66	1	118.14	123.74
X <sub>1</sub> (X <sub>3</sub> X <sub>6</sub> X <sub>2</sub> X <sub>4</sub> )	0.49	6.928	5.01	1	121.43	129.86	X <sub>1</sub> X <sub>3</sub>	1.51	6.811	1.11	1	118.00	122.21
X <sub>1</sub> X <sub>3</sub> X <sub>6</sub> X <sub>2</sub> X <sub>4</sub> X <sub>5</sub>	0.26	7.068	1.00	1	123.36	133.17	X <sub>1</sub>	9.74	6.993	1.41	1	118.63	121.43

기준 1: |t| < t<sub>0.05(n-p)</sub> 절단  $\Rightarrow X_1, X_3$  종결

min|t| > 1 인곳에서 차례대로 X<sub>1</sub>, X<sub>3</sub>, X<sub>6</sub> 종결.

기준 2: |t| < 1 절단  $\Rightarrow X_1, X_3, X_6$  종결. \* RMS가  $t_{0.05}$ 로 봄을 따라  $t_{0.05}$ 로 봄을 따라 차이.

### • 단계적 방법

FS와 비슷하지만 t<sub>0.05</sub> 추가. t<sub>0.05</sub>에서 삭제를 한다.

### C<sub>p</sub>

앞의 두 가지 방법과 유사성이 다르다. (RMS, Rank 등)

$$C_p = \frac{SSE}{n^2} + (2p-n)$$

~~→ 고른 방법 (앞에 고른 C<sub>p</sub>보다 커야 한다, 그 그림의 P값과 가장 비슷한 값을 고른다)~~

좋은 C<sub>p</sub>값을 얻기 위해서는 T<sup>2</sup>에 대한 좋은 결정값을 얻는 것이다.

\* 비공선성에서 FS보다 BE가 루틴되고

공선성에서도 FS보다 BE가 더 추천된다.

\* RMS 기준으로 변수를 설정할 경우.  $X_1 X_4 X_6$  의 RMS 보다 큰 full 모형은 적합지 않고 볼 수 있다.

\* AIC와 BIC 기준으로 선택한다면 AIC는  $X_1 X_3$ , BIC는  $X_1$  라고 할 수 있다. 근데 그 차이가 나는 모형들이 후보군이 된다고 했으나 AIC는  $(X_1, X_2, X_3, X_4, X_5, X_6)$  424, BIC는  $(X_1, X_3)$  272가 선택된다.

두 개가 차이나는 이유는 파생변수 때문의 특성으로 BIC는 변수가 3개를 선정하는 때문이다.

$$AIC = n \log\left(\frac{SSE}{n}\right) + 2p \quad BIC = n \log\left(\frac{SSE}{n}\right) + p \log n$$

예) 실증적 테이터.

$VIF_1, VIF_3 > 10$  이로 공선성이 있음을 알 수 있다.

변수	(a)	(b)	(c)	(d)	(e)	(f)	(g)
G계수	0.96			1.15	0.87	0.24	
t검정	11.10			11.90	1.62	0.68	
M계수		0.55		-0.21		-0.40	-0.43
t검정		2.16		-2.19		-4.41	-5.35
W계수			-0.95		-0.09	-1.02	-1.28
t검정			-9.77		-0.17	-2.11	-15.90
$R^2_a$	0.91	0.24	0.89	0.95	0.90	0.91	0.97

\* FS 선택

$$1단계 \quad Y = \beta_0 + \varepsilon$$

$$2단계 \quad Y = \beta_0 + \beta_1 G + \varepsilon \quad (\text{선택})$$

$$3단계 \quad Y = \beta_0 + \beta_1 G + \beta_2 M + \varepsilon \quad (1\text{st} \text{剔除})$$

$$4단계 \quad Y = \beta_0 + \beta_1 G + \beta_2 M + \beta_3 W + \varepsilon \quad (R^2_a \text{ 증가})$$

근데 (d)-(f) 구간에서 두의 유의성이 큰 변수가 생겼으므로

공선성을 의심해본다.

\* BE 선택

$$1단계 \quad Y = \beta_0 + \beta_1 G + \beta_2 M + \beta_3 W + \varepsilon$$

$$2단계 \quad Y = \beta_0 + \beta_2 M + \beta_3 W + \varepsilon \quad (\text{선택})$$

$$3단계 \quad 모든 \beta_i \text{의 값이 } 1보다 크므로 중단.$$

$\Rightarrow$  FS와 BE의 G 변수가 가지는 성질이 다른 이유는 공선성 때문이다. 2단계에서  $VIF_3$  가 크고  $R^2_a = 15.6 > 15$  이므로 공선성이 존재하는 것을 알 수 있다.

$\Rightarrow$  공선성이 있더라도 공선성이 시계열로 오차항의 자기상관성이 있는 자료 등 여러 문제점이 있을 수 있으므로 드물게 기계적인 변수선택은 하면 안된다.