제 5 장 다중회귀분석 (Multiple Regression Analysis)

1. 다중회귀 모형의 구조

1) 다중회귀분석이란: 설명변수(독립변수)가 2 개 이상인 회귀모형을 분석대상으로 삼고있다

→ 기본가정: 설명변수는 2 개이며, 각설명변수는 종속변수와 선형관계에 있다

- 다중분석의 의의: 분석내용을 향상시킬 수 있다

a) 추가적인 독립변수를 도입함으로써 오차항의 값을 줄일 수 있다

b) 단순회귀분석의 단점을 극복 할 수 있다

→ 종속변수를 설명하는 독립변수가 두개일 때 단순회귀모형을 설정한다면 모형설정(specification)이 부정확할 뿐 아니라 종속변수에 대한 중요한 설명변수(독립변수)를 누락함으로써 계수 추정량에 대해 편의(bias)를 야기 시킬 수 있으므로 단순회귀분석은 그 유용성을 상실하게 된다. 따라서 다중회귀분석을 통해 편의현상(bias)을 제거할 수 있다

2) 다중회귀모형의 일반형

- 종속변수 Y_i 가 상수항과 k 개의 독립변수 X_{1i}, X_{2i} .. X_{ki} 에 의해 설명되는 모집단의 다중회귀모형은, 기본모형: $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i$

  → 설명변수 k 개, 모수 (k+1)개

- Matrix 형태

3) 다중회귀 모형의 기본가정

a) 회귀모형은 모수에 대해 선형인 모형이다: $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$

b) 독립변수 X_{1i} , X_{2i} 는 비확률(nonstochastic)이다

c) 오차항의 평균은 0 이다: $E(\varepsilon_i) = 0$

d) 오차항의 분산은 모든 관찰치에 대해 σ^2 의 일정한 분산을 갖는다 (동분산: homoskedasity): $Var(\varepsilon_i) = E[\varepsilon_i - E(\varepsilon_i)^2] = E(\varepsilon_i^2) = \sigma^2$

e) 서로다른 관찰치간의 오차항은 상관이 없다: $Cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i\varepsilon_j) = 0$ $(i \neq j)$

  → 오차항들은 서로 독립적이며 그들의 공분산은 0 이다

f) 오차항은 각독립변수와 독립적이다: $E(X_i, \varepsilon_i) = X_iE(\varepsilon_i) = 0$

g) 오차항이 정규분포를 따르며, $\varepsilon_i \sim N(0,\sigma^2)$

→ 이상의 가정들은 기존의 단순회귀분석의 경우와 기본적인 내용에 있어서 큰 차이가 없으나 다음의 두가지 가정은 다중회귀분석에만 적용이 된다

h) 독립변수간에는 정확한 선형관계가 없다: $\rho(X_{1i}, X_{2i}) \neq \pm 1$

→ 한 독립변수가 다른 독립변수와 1 차함수관계에 있어서는 안된다

→ 만일 $\rho(X_{1i}, X_{2i})$의 절대값이 1 에 가까워지면, 최소자승법의 적용이 어렵게 되며, 이러한 현상을 다중공선성(multicollinearity)이라 하고 특히, $\rho(X_{1i}, X_{2i}) = \pm 1$ 인 경우에는 완전공선성(perfect collinearity) 이라 한다.

i) 관측된 값들의 수는 독립변수의 수보다 최소한 2 이상 커야 한다.

→ 독립변수의 수가 k 라면, 미지수인 절편을 포함하여 총(k+1)의 모수를 추정해야하며, 이를 위해 관측치 수(표본수:n)은 (k+1)보다 커야 연립방정식 체계를 이용하여 최소자승법을 사용할 수 있다: $n > k+1$

→ 자유도는 n-(k+1)이고, 최소한 자유도가 1 이상 되기위해 $n > k+2$ 가 필요하다

4) 다중회귀분석 계수의 해석

- 독립변수가 여러 개인 다중회귀분석에서의 회귀계수의 해석은 다른 독립변수가 불변일 때 (통제된 상태에서), 해당되는 독립변수의 변화에 따른 종속변수의 평균변화량을 나타내는 직접효과(direct effect) 또는 순효과(net effect)를 뜻한다.

  → 이는 독립변수가 여러 개이므로 이들간에 서로 상관관계를 있을 수 있기 때문에 한 독립변수의 값이 다른 독립변수에 영향을 미칠 수 있다

2. 최소자승법에 의한 다중회귀모형의 추정

- 모집단에 대한 기본가정들이 충족된다는 가정하에 최소자승법을 이용하여 표본회귀선을 도출할 수 있다

→ 다중회귀모형의 표본회귀 모형:$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$

  ⇨ 모집단의 모수인 α, β_1, β_2 에 대한 최소자승 추정량 ($\hat{\alpha}$, $\hat{\beta}_1$, $\hat{\beta}_2$)은 최소자승법을 이용하여 추정치를 구한다 (표본의 크기: n)

  ⇨ $\min L = \min [\sum_{i=1,n} e_i^2] = \min [\sum_{i=1,n} (Y_i - \hat{Y}_i)^2] = \min [\sum_{i=1,n} (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})^2]$,

  ⇨ $(\partial L/\partial \hat{\alpha}) = (\partial L/\partial \hat{\beta}_1) = (\partial L/\partial \beta \hat{\beta}_2) = 0$

⇨ 일반적으로 추정계산과정이 매우 복잡하고 시간이 많이 소요되는 관계로 컴퓨터 프로그램을 이용하여 계산한다.

⇨ 최소자승 추정량의 특성

a) 불편의 추정량 (unbiased estimator): $E(\hat{\beta}_i) = \beta_i$

b) 일치 추정량 (consistent estimator): $\hat{\beta}_i \rightarrow \beta_i$ as $n \rightarrow \infty$

c) BLUE (Gauss-Markov Theorem)

→ 추정량의 분산

a) $var(\hat{\beta}_1) = \sigma^2/[(1 - r_{23}^2)(\sum(X_{2i} - \mathbf{X_2})^2)]$, b) $var(\hat{\beta}_2) = \sigma^2/[(1 - r_{23}^2)(\sum(X_{3i} - \mathbf{X_3})^2)]$

➔ $r_{23} = [\sum (X_{2i} - \mathbf{X_2})^2(X_{3i} - \mathbf{X_3})^2]/[\sqrt{\sum}(X_{2i} - \mathbf{X_2})^2 \sqrt{}(X_{3i} - \mathbf{X_3})^2]$

⇨ 추정량의 분산에 영향을 미치는 요인들

a) 오차항의 분산(σ^2)이 작을수록 추정량의 분산이 작아진다

b) 표본의 크기(n)가 클수록 추정량의 분산이 작아진다

c) 설명변수의 편차($X_{ki} - \mathbf{X_k}$))들이 클수록 추정량의 분산이 작아진다

d) 설명변수들간의 상관관계(r_{23})가 작을수록 추정량의 분산이 작아진다

⇨ 설명변수들간의 상관관계(r_{23})가 클수록, 설명변수들간의 공선성

(collinearity)이 강해지므로 추정량의 분산값이 커지게 되어 정확한 추정량의 값을 구하기가 어렵게 됨.

- 오차항의 분산 추정치: 오차항의 분산(σ^2)이 일반적으로 알려져있지 않으므로 표본의 분산(s^2)을 대신 이용한다 ➔ $s^2 = \sum_{i=1,n} e_i^2 /(n-(k+1))$ (k 는 독립변수의 개수)

3. 다중회귀모형에 대한 가설검정

- 다중분석회귀의 기본가정이 성립되면 최소자승 추정량은 BLUE 특성을 가지나 각 독립변수가 종속변수의 변화를 설명하는데 유용한지의 여부를 검정하여야 한다

1) 모수 β 에대한 가설검정: t-검정

- 개별 회귀계수 β 에 대한 가설검정은 t-분포를 이용하여 이루어진다

→ $(\alpha\hat{} - \alpha) / s_{\alpha\hat{}} \sim t(n-k-1)$, $(\beta\hat{}_j - \beta_j) / s_{\beta\hat{}j} \sim t(n-k-1)$, $j = 1,2$.

⇨ $\varepsilon_i \sim N(0,\sigma^2)$ 가정하에서 추정량($\alpha\hat{}$ $\beta\hat{}_1$ $\beta\hat{}_2$)들이 정규분포를 따르는 오차항 ε_i 와 1 차함수관계를 갖게 되므로 이들 역시 정규분포를 따르게 되고, 모분산값, $var(\varepsilon_i) = \sigma^2$ 이 알려지지 않았기 때문이다.

⇨ 이 경우 임계치 (critical value)를 구할 때 자유도가 (n-k-1)임을 유의

- 귀무가설 H_0: $\beta\hat{}_j = \beta^0_j$ 의 가설을 검증하는 경우 단순회귀분석의 경우와 같이 양측검정, 단측검정의 방법이 사용되며, 2-t 으뜸법칙과 유의확률에 의한 가설검정을 이용

2) 모형에 대한 검정: F-검정

- χ^2 분포와 F-분포

a) χ^2 분포: 확률변수 X_1, X_2,… X_n 이 정규분포 $N(\mu,\sigma^2)$을 따를 때, 표준화변수 $(X_i-\mathbf{X})$의 제곱합의 분포는 자유도(n)을 가진 χ^2 분포를 따른다

→ $\Sigma(X_i-\mathbf{X})^2/\ \sigma^2 \sim \chi^2(n)$

→ F-분포는 (0,∞)의 구간에서 긴오른쪽 꼬리를 가진 형태를 가진다.

b) F-분포: 확률변수 X_1, X_2,… X_{m1} 이 정규분포 $N(\mu_1,\sigma_1^2)$을 따르고, Y_1, Y_2,… Y_{m2} 이 정규분포 $N(\mu_2,\sigma_2^2)$을 따를 경우, 표준화변수의 제곱합의 비율은 자유도(m1,m2)를 가진 F 분포를 따른다.

→ $\{[\Sigma(X_i-\mathbf{X})^2/\sigma_1^2]/m1\}/\{\ [\Sigma(Y_i-\mathbf{Y})^2/\sigma_2^2]/m2\} \sim F(m1,\ m2)$

[m1:분자 자유도(numerator degree of freedom: df_n), m2: 분모자유도 (denominator degree of freedom: df_d)]

→ F-분포는 (0,∞)의 구간에서 긴오른쪽 꼬리를 가진 형태를 가진다.

⇨ 만일 확률변수 $T\sim t(m2)$이면, $T^2\sim F(1,m2)$의 분포를 갖는다

⇨F-분포 형태

- 모형의 적합도(유의성)에 대한 가설검정

→ t-분포를 이용한 검정은 단순히 개별 독립변수들이 종속변수의 변화를 설명하는데 유용한지의 여부를 확인할 수 있으나 모든 독립변수집단 전체가 총체적으로 보아 종속변수를 설명할 수 있는지의 여부를 확인 할 수 없다

→ 모형의 적합도에 대한 가설검정은 모든 독립변수가 종속변수에 영향을 미치는가를 검정한 것이다

→ 두개의 독립변수를 가진 모형인 경우: $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$

⇨ 모형의 유의성 검정

⇨ H_0: $\beta_1 = \beta_2 = 0$ vs. H_1: 적어도 하나의 β 는 0 이 아니다

a) 평균오차자승 합(MSE: mean square error) = RSS / (n-k-1), RSS = $\sum_{i=1,n} e_i^2$

b) 평균회귀자승 합(MSR: mean square regression) = ESS/k, ESS = $\sum_{i=1,n} (Y\hat{} - Y)^2$,

➔ $F_{(k,\ n-k-1)}$= MSR/MSE = $[\sum_{i=1,n} (Y\hat{} - Y)^2 /k] / [\sum_{i=1,n} e_i^2 /(n-k-1)]$ ~ F(k,n-k-1): 만일 F-통계치가 유의수준 α 에서 자유도(k-1, n-k)의 임계치(F-분포값) 보다 크면, 귀무가설이 기각되어 독립변수 전체가 종속변수의 설명에 유의적 (statically significant) 따라서 독립변수 전체가 종속변수 설명에 유의

적이지 못하다(statically insignificant) .

- 회귀계수의 일부분에 대한 가설검정: F-검정법

→ 가설들에 대한 F-검정법은 제약없는 원래의 모형(unrestricted model)으로부터 도출된 잔차의 제곱합(RSS_U)과 귀무가설이 참이라는 가정하에서의 모형(restricted model)에서 도출된 잔차의 제곱합(RSS_R)을 비교하는 것에 기반을 둔 검정방법이다

→ 만일 귀무가설에서 가설의 수(제약식의 수)가 J 개일 경우,

$(RSS_R - RSS_U)/\sigma^2 \sim \chi^2(J)$ 그리고 $(RSS_U)/\sigma^2 \sim \chi^2(n-k-1)$의 분포특성을 갖는다.

⇨ F-검정통계치 = $[(RSS_R - RSS_U)/J] / [(RSS_U)/(n-k-1)] \sim F(J,n-k-1)$

⇨ 만일 F-검정통계치 값이 주어진 유의수준에 의한 임계치값, 자유도 (J,n-k-1)의 F-분포값 보다 클 경우 귀무가설을 기각하게 된다

⇨ 만일 F-검정통계치 값이 주어진 유의수준에 의한 임계치값(F_c), 자유도(J,n-k-1)의 F-분포값 보다 작을 경우 귀무가설을 채택하게 된다

→ 독립변수가 k 개인 다중회귀모형 $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + .. + \beta_k X_{ki} + e_i$ 에서 회귀계수 β 의 일부분에 대한 가설검정을 한다

→ 제약된 최소자승 추정치(restricted least squares estimate)와 제약되지않은 최소자승 추정치(unrestricted least squares estimate) 사용 한다

→ 예) H_0: $\beta_{k-1} = \beta_k = 0$ (X_{k-1}와 X_k의 두변수만 종속변수에 영향이 없다)

a) 제약된 최소자승 추정치(restricted least squares estimate): 제약조건을 고려한 최소자승 추정치를 구하는 방법

⇨ 귀무가설의 조건을 고려한 회귀분석모형은 $Y_i = \alpha + \beta_1 X_{1i} + ... + \beta_{k-2} X_{k-2i} + e_i$

⇨ 이 제약된 모형을 이용하여 Y 와 X_1 ..X_{k-2} 에 대하여 회귀분석하여 잔차항의 지곱합(restricted RSS: $RSS_R = \sum_{i=1,n} e_i^{*2}$)를 구한다

⇨ 귀무가설의 제약조건을 감안하지않은 일반적인 회귀모형, $Y_i = \alpha + \beta_1 X_{1i} + ... + \beta_k X_{ki} + e_i$ 을 이용하여 Y 와 X_1 ..X_k 에 대하여 회귀분석하여 잔차항의 지곱합(unrestricted RSS: $RSS_u = \sum_{i=1,n} e_i^2$)를 구한다

⇨ 만일 귀무가설이 맞다면, $RSS_R \approx RSS_u$ 의 관계가 성립하고, 귀무가설이 성립하지 않으면 $RSS_R > RSS_u$ 이 관계가 되어 두 값간에 차이가 생겨난다.

⇨ 이 경우 가설검정 역시 F-분포를 이용하여 실행한다:

$F = [(RSS_R - RSS_u)/J] / [RSS_u /(n-k)] \sim F(J,n-k-1)$, J 는 제약조건의 계수

- 회귀계수의 선형결합에 대한 검정

- 독립변수가 k 개인 다중회귀모형 $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + .. + \beta_k X_{ki} + e_i$ 에서 회귀계수 β 의 선형결합에 대한 귀무가설을 검정한다

- 예) H_0: $\beta_1 + \beta_2 = 1$

→ 귀무가설하에서 $\beta_1 = 1 - \beta_2$ 이 되며, 이 조건하에서 제약된 회귀모형은

$Y_i = \alpha + (1-\beta_2)X_{1i} + \beta_2 X_{2i} + .. + \beta_k X_{ki} + e_i$ 이 되며 , 이는 다시

$Y_i - X_{1i} = \alpha + \beta_2(X_{2i} - X_{1i}) + .. + \beta_k X_{ki} + e_i$ 로 바꿀 수 있다

⇨ 이 조건하에서 종속변수($Y - X_1$)를 독립변수($X_2 - X_1$), … X_k 에 대해 회귀분석하여 제약된 모형에서의 잔차항의 제곱합 ($RSS_R = \sum_{i=1,n} e_i^{*2}$)을 구한다: 이때, β_1 의 추정치는 $1- \hat{\beta}_2$

⇨ 제약조건을 감안하지 않은 일반적인 모형으로부터 잔차항의 제곱합 ($RSS_u = \sum_{i=1,n} e_i^2$)을 구한다

⇨ 검정통계량, $F = [(RSS_R - RSS_u)/J] / [RSS_u /(n-k)] \sim F(J,n-k-1)$, J 는 제약식 수

4. 결정계수(R^2)와 조정된 결정계수(Adjusted R^2: R^2): 적합도 검정

- 단순회귀분석의 경우와 같이 결정계수 도출에 필요한 관계식을 정리해보면,

  → $\sum_{i=1,n} (Yi - Y)^2 = \sum_{i=1,n} (Y\hat{}i - Y)^2 + \sum_{i=1,n} (Yi - Y\hat{}i)^2$: TSS = ESS + RSS

  → 결정계수, R^2 = ESS/TSS = 1- (RSS/TSS)

- 다중회귀분석의 경우 독립변수가 여럿 개이기 때문에, R^2 의 값이 독립변수의 수가 많아짐에 따라 증가한다.

  → 독립변수의 수가 증가함에 따라 잔차항의 제곱 합($\sum_{i=1,n} e_i^2$)이 감소하게 되고 결과적으로 R^2 의 값이 증가하게 된다.

  → 이는 R^2 의 값을 정의함에 있어 단순히 전체 변동량과 회귀선에 의해 설명되는 변동량을 비교함으로써 각 변동량이 가지는 자유도를 고려하지 않았기 때문이다.

  → 단순히 R^2 의 값을 증가시키기 위해 종속변수의 설명에 중요하지 않은 독립변수를 추가 시키는 경우가 발생할 수 있다

- 이러한 문제점을 방지하고 보다 적합한 평가기준을 위해 자유도를 동시에 고려하는 조정된 결정계수(Adjusted R^2)를 사용한다

  → 조정된 결정계수(Adjusted R^2)를 도출할 경우, 단순한 변동량 대신 자유

도 개념을 구체적으로 고려대상에 포함 시킬 수 있는 분산개념을 사용한다

Adjusted R^2 = 1- $[\sum_{i=1,n} e_i^2/(n-k-1)] / [\sum_{i=1,n} (Yi - Y)^2/(n-1)]$ = 1 – (1- R^2)[(n-1)/(n-k-1)]

⇨ k>1 이면, $R^2 > R^2$

⇨ 다른 조건이 일정할 때 독립변수 k 가 증가함에 따라 R^2 값은 감소한다: 종속변수의 설명에 유의적(significant)이지 않거나 중요하지않은 독립변수를 포함시켜 결정계수를 증가 시키는 것을 방지 할 수 있게 한다

⇨ $R^2 > 0$ 이나 $R^2 < 0$ 이 가능하다: 이는 R^2 값이 아주 작은 경우로 고려하고 있는 모형이 자료에 적합하지 않음을 의미한다

⇨ 종속변수가 서로다른 두 모형을 비교할 때, 결정계수값을 비교하는 것은 의미가 없다: 이는 종속변수의 전체변동값이 다르기 때문에 결정계수값을 정확히 비교할 수가 없다

5. 예측

- 기본가정이 충족된다고 볼 수 있는 상황에서 최소자승법에 의해 도출된 표본회귀선이 적합도나 통계적 유의성 검정에 의한 평가 결과 합당한 추정식으로

판정이 되면, 이 추정식을 이용하여 신뢰할 수 있는 예측을 시행할 수 있다.

- 예측방법은 단순회귀분석과 동일하다:

→ 점예측: 도출된 다중표본회귀식에 임의로 내정된 각독립변수의 값을 대입함으로써 성립된다: $\hat{Y}_f = \hat{\alpha} + \hat{\beta}_1 X_{1f} + \hat{\beta}_2 X_{2f}$

a) X_{1f} 와 X_{2f} 는 회귀선의 추정과정에서 사용되지 않은 추가적인 값들이며 분석자의 판단에 의거 적절히 책정되는 것이 일반적이다

b) 표본회귀선에 의해 표출된 $\hat{Y}_f$ 란 독립변수의 값(X_{1f}, X_{2f})이 주어졌을 때 평균적으로 기대할 수 있는 종속변수값으로 예측대상이 긴 미래에 예측하거나 예측에 사용된 독립변수의 값들이 표본회귀선 도출에 사용되었던 독립변수값들의 평균값과 많은 차이가 있을 때, 점예측치의 실현가능성은 희박해진다.

c) 점예측은 가능한 가까운 미래예측에 한정하여 사용하는 것이 타당하다