4.3 중회귀분석의 응용

1) 실생활에서의 중회귀분석 사례
2) 총변동의 분해

1. 실생활에서의 중회귀분석 사례

2000년 서울 강남지역의 아파트 가격

(추정된 아파트 가격)= -20.394+1,549(평수)+0.76(연령)　단위: 만 원

- (i) 아파트 평수와 (ii) 아파트 연령의 2개 설명변수 갖는 중회귀분석 모형임
- 하지만 여전히 중요한 제3의 변수가 모형에서 누락된 결과 “아파트가 red wine처럼 오래될수록 비싼” 것으로 잘못 추정됨
- 아파트 단지규모는 아파트 연령과도 관련되어 있고(강남 개발 초기에 지어진 아파트가 대단지로 들어섬) 동시에 아파트 가격과도 밀접하게 연관되어(대단지는 편의시설이 발달되어 있어 아파트 값이 비쌈) 있는 혼동요인(confounding factor)으로 작용하고 있음

1. 실생활에서의 중회귀분석 사례

2000년 서울 강남지역의 아파트 가격

(추정된 아파트 가격)= -20.291+1,538(평수)-137(연령)+2(단지규모)

- 위 식은 (i) 아파트 평수, (ii) 아파트 연령 등 기존 2개의 설명변수에 (iii) 아파트 단지규모라는 설명변수를 추가하여 확장한 중회귀분석 모형을 추정한 식임
- 제3의 요인인 아파트 단지규모를 통제한 결과 아파트 연령과 가격간의 관계가 보다 상식에 부합되게 얻어짐→동일한 (평수, 단지규모)에 속하는 아파트를 비교해 보면 오래된 아파트가 1년당 평균 137만원 꼴로 값이 낮음
- 물론 오래된 아파트의 재건축 가능성을 고려하면 아파트 연령과 아파트 가격간의 관계는 비선형일 수도 있을 것으로 예상됨

1. 실생활에서의 중회귀분석 사례

회귀분석: 홈런 on BB

- **규정타석을 채운 타자들로만 이루어진 "동질적" 자료 이용**
- 홈런이 많은 타자는 큰 것 '한 방'을 노리기 때문에 삼진도 많이 당함
  - Babe Ruth: "I hit big, I miss big. I would like to live as big as I could."

$$(\text{사사구수}) = \alpha + 0.51(\text{홈런수}) + \epsilon$$

1. 실생활에서의 중회귀분석 사례

회귀분석: 홈런 on BB

- **모든 타자들로 이루어진 “이질적” 자료 이용 (타석수 통제되지 못함)**
- 홈런이 많은 타자는 큰 것 ‘한 방’을 노리기 때문에 삼진도 많이 당함

$$(\text{삼진수}) = \alpha + \beta(\text{홈런수}) + \epsilon$$

$$b = 2.40 \ (\gg 0.51)$$

1. 실생활에서의 중회귀분석 사례

중회귀분석: 홈런 on (BB and 타석수)

- **모든 타자들로 이루어진 “이질적” 자료 이용하되 중회귀분석 이용하여 통계적 방법으로 타석수 통제**
- 홈런이 많은 타자는 큰 것 ‘한 방’을 노리기 때문에 삼진도 많이 당함

$$(\text{삼진수}) = \alpha + \beta_1(\text{홈런수}) + \beta_2(\text{타석수}) + \epsilon$$

$$b_1 = 0.63 \qquad b_2 = 0.14$$

1. 실생활에서의 중회귀분석 사례

결혼시장 분석

- 심리학, 사회학, 인류학 등에서 행하는 설문조사를 통한 남녀 배우자 선호 비교 연구: 정직한 진술 (truth telling)이 보장되지 않는 맹점이 있음
- 실제 선택한 결과를 관측한 현시선호(revealed preference)의 데이터를 연구할 필요가 있음.
- 국내 모 결혼정보회사의 상세한 개인 프로필 및 선택에 대한 현시 선호 데이터를 사용하여, 우리 나라 중매결혼시장에서 남녀의 배우자 선호의 차이를 비교함. **사회 경제적인 조건과 외모 조건에 대한 선호에 있어서 남녀의 차이가 어떻게 드러나는가?**

1. 실생활에서의 중회귀분석 사례

결혼시장 분석 : 사회경제적 조건 및 신체적 조건들

	남			여		
특성	중간값	평균	표준편차	중간값	평균	표준편차
나이	32.2	33.4	3.4	29.6	30.3	2.7
신장(cm)	173	173.7	4.3	162	162.7	3.9
체중(kg)	68	69.3	6.8	50	50.2	4.1
인상등급(0~5)	3	3	1.2	3	3	1.1
연봉(원)	3500	4833	2575	2200	3504	1860

학력	남	여
대학원 이상(%)	22	28
대졸 이상(%)	89	83
고졸(%)	11	17

결혼 적령기의 전체 인구집단과 비교할 때, 결혼정보회사의 회원들은 나이는 다소 많고, 신장은 다소 크고 연봉은 많으며, 학력이 높음.

1. 실생활에서의 중회귀분석 사례

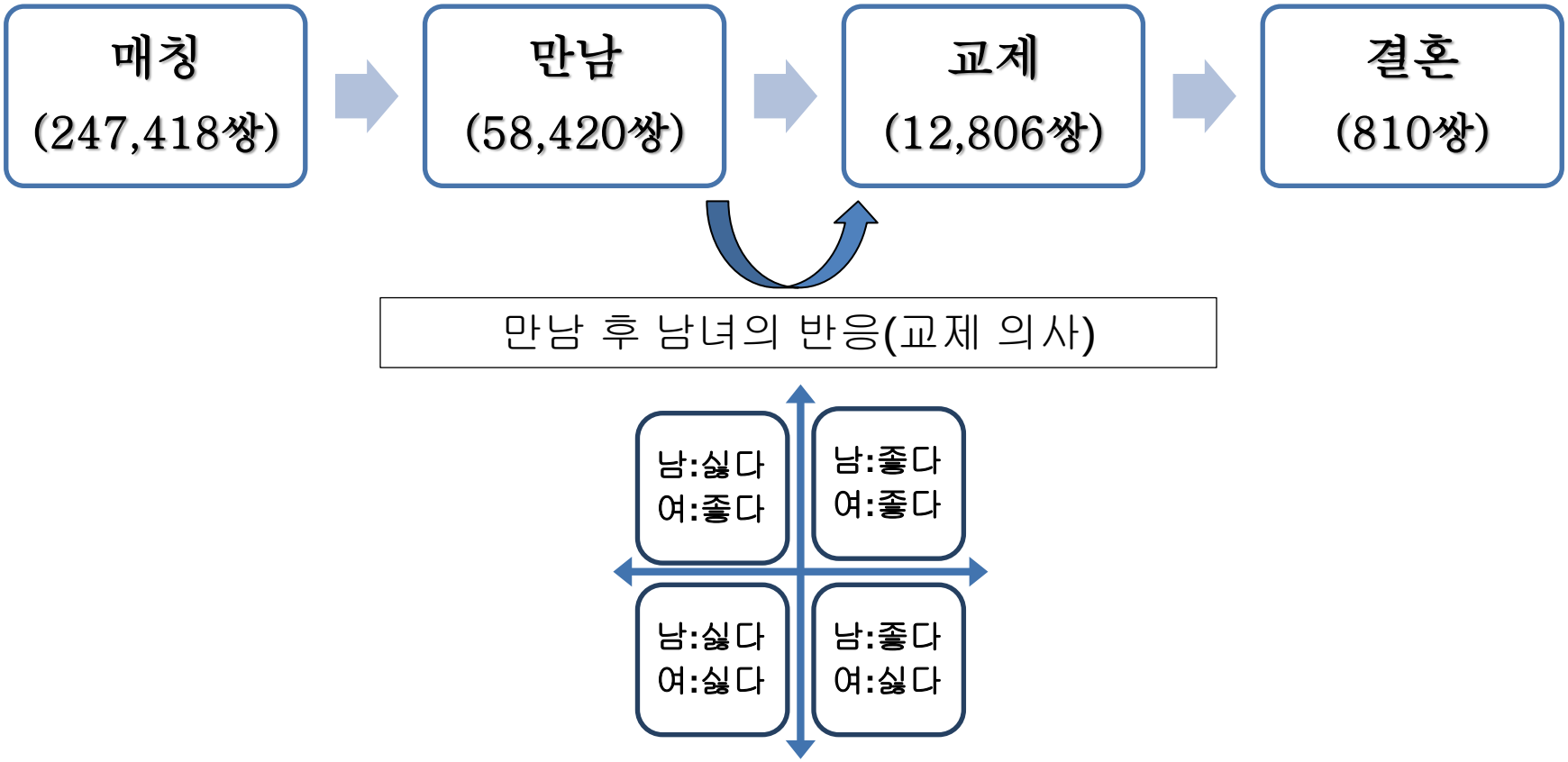
결혼시장 분석:'결혼정보회사'의 배우자지수

- 사회경제적 위세 지수 (SESI: Socio Economic Status Index)
  - 학력, 학벌, 직업, 소득 등을 포괄하는 지수 (회사 측에서 작성).
- 신체적 매력 지수(PAI: Physical Attractiveness Index)
  - 키, 체중, 인상등급 등을 포괄하는 지수 (회사 측에서 작성).
- 가정환경 지수 (FBI: Family Background Index)
  - 부의 학력, 부의 직업, 부의 재산, 양친 생존여부, 부모 이혼여부, 형제관계 등을 포괄하는 지수로 회사에서 작성.

지수	남					여				
	최저값	중간값	최고값	평균	표준편차	최저값	중간값	최고값	평균	표준편차
SESI	34.1	73	100	72	11	33.2	71	99.7	70	13
PAI	21	80	100	76	12	14	82	100	79	10
FBI	7.6	54	99.3	54	19	7.6	62	98.1	61	28

1. 실생활에서의 중회귀분석 사례

결혼시장 분석 : 결혼에 이르는 단계와 만남 후 남녀의 반응



1. 실생활에서의 중회귀분석 사례

결혼시장 분석 : 분석모형 및 추정 결과

- (반응) $= \alpha + \beta_1$(상대의 $SESI$) $+ \beta_2$(상대의 PAI) $+ \beta_3$(상대의 FBI) $+ \epsilon$
- 반응: 좋다=1, 싫다=0

	남자의 반응		여자의 반응	
	추정치	표준오차	추정치	표준오차
SESI	0.0123	0.0130	0.0191*	0.0014
PAI	0.0319*	0.0034	0.0118*	0.0027
FBI	0.0139*	0.0053	0.0029	0.0087

1. 실생활에서의 중회귀분석 사례

결혼시장 분석 : 해석과 함의

사회경제적 조건과 외모의 교환 관계(Trade off between SESI & PAI)

- 남자가 여자를 평가할 때는 사회경제적 조건(SESI)에 비해 외모(PAI) 중시
- 여자가 남자를 평가할 때는 외모(PAI)에 비해 사회경제적 조건(SESI) 중시

2. 총변동의 분해

통화증가율과 인플레이션율

표 8-2 우리나라의 연간 통화증가율과 인플레이션율, 1986-2008년도 (단위: %)

연도	통화증가율	인플레이션율	연도	통화증가율	인플레이션율	연도	통화증가율	인플레이션율
1986	26.4	0.88	1994	21.85	5.2	2002	13.93	3.81
1987	34.79	6.22	1995	21.1	4.81	2003	2.36	3.39
1988	28.3	6.78	1996	18.02	4.68	2004	6.33	3.4
1989	25.78	7.32	1997	21.38	8.29	2005	7.27	2.22
1990	25.31	9.16	1998	21.27	1.46	2006	11.3	1.68
1991	20.13	7.72	1999	3.19	1.89	2007	12.47	3.89
1992	21.95	4.51	2000	5.16	3.44	2008	11.96	3.75
1993	17.54	6.35	2001	8.53	2.64			

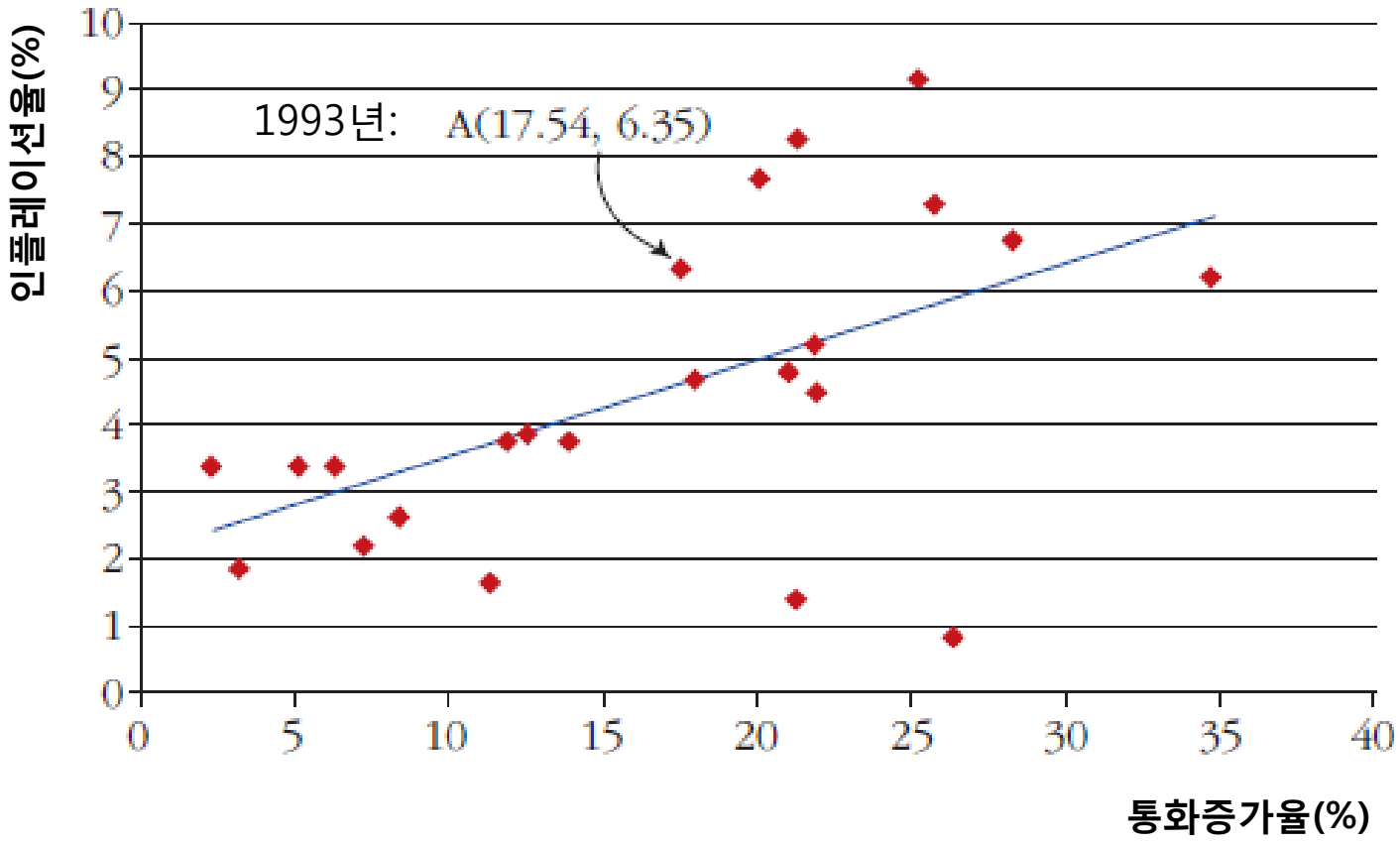
주: 통화증가율은 M2증가율이고 인플레이션율은 소비자물가지수 상승률임.

류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 175

- 1986~2008년간의 평균 인플레이션율은 4.50%였는데 반해 1993년의 인플레이션율은 6.35%로 표본 내 23년간의 평균보다 1.85% 높음
- 이중 얼마만큼의 차이를 통화증가율의 차이로 설명할 수 있을까?

2. 총변동의 분해

통화증가율과 인플레이션율



류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 175

2. 총변동의 분해

통화증가율과 인플레이션율

- 1993년: $(\boldsymbol{x_i}, \boldsymbol{y_i}) = (17.54\%, 6.35\%)$

$$y_i - \bar{y} = [(a + bx_i) - \bar{y}] + [y_i - (a + bx_i)]$$
$$T \quad = \quad R \quad + \quad E$$

- T : 1993년의 인플레이션율$(\boldsymbol{y_i})$이 지난 23년의 평균과 차이나는 부분 전체
- R : 그 전체 가운데 1993년의 통화증가율$(\boldsymbol{x_i})$로 설명되는 부분
- E : 그 전체 가운데 1993년의 통화증가율로는 설명되지 않는 부분

2. 총변동의 분해

총변동의 분해

$$\sum (y_i - \bar{y})^2 = \sum [(a + bx_i) - \bar{y}\,]^2 + \sum [y_i - (a + bx_i)]^2$$

SST = SSR + SSE

(회귀식으로 설명됨) (회귀식으로 설명 안 됨)

- **SST**[총제곱합 (total sum of squares)]: y의 평균 주위로의 총변동
- **SSR**[회귀제곱합 (regression sum of squares)]: 회귀직선에 의해 설명되는 변동분
- **SSE[**잔차제곱합 (residual sum of squares) 또는 오차제곱합 (error sum of squares)]: 회귀직선에 의해 설명되지 않는 변동분

2. 총변동의 분해

결정계수(R^2)

- 결정계수(R^2) = 총변동에서 차지하는 설명되는 변동분의 비율

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad (0 \leq R^2 \leq 1)$$

- 결정계수의 값이 1에 가까울수록 회귀직선의 설명력은 높다.
- 단순회귀분석의 경우 결정계수인 R^2 값은 두 변수간 상관계수인 r 의 제곱과 같게 된다. (단순회귀분석의 경우에는 R^2 = r 제곱)

2. 총변동의 분해

조정된 결정계수(adjusted R^2)

- 설명변수를 추가하면 추가할수록 R^2는 언제나 증가함
  - R^2=1-SSE/SST 인데 SST는 고정된 반면 SSE는 설명변수 추가될수록 감소
- 이 문제를 해결하기 위해 아래의 "조정된 결정계수"를 정의함

$$\overline{R}^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} \quad (n = \text{표본크기},\ k = \text{설명변수의 개수})$$

  - SSE와 SST가 각각의 자유도로 나누어진 형태로 등장
  - SST의 자유도=(n-1): 표준편차 구할 때의 자유도와 동일
  - SSE의 자유도=(n-k-1): n개 자료 이용 총 (k+1)개의 계수 추정한 결과
  - 조정된 결정계수는 설명변수가 추가된다고 해서 반드시 늘지는 않음