Chapter 16

Regression Analysis II

경영대학 재무금융학과

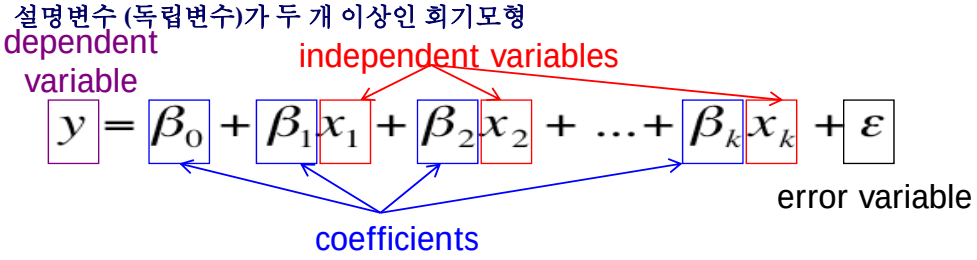
윤선중

Objectives

- 다중회귀모형 (multiple regression model)
  - 단순회귀모형의 확장
- 회귀 모형의 평가 : 3단계
  - 추정치의 표준오차
  - 결정계수(자유도 조정결정계수)
  - 분산분석형태의 F검정
- 독립-종속 변수간의 선형관계 결정을 위한 t-검정사용 가능
- 다중공선성 (multicollinearity)
- 제 1계 자기상관 (first-order authcorrelation)
- 더빈-왓슨 검정 (Durbin-Watson test)

Multiple Regression Model

- 정의
  - 설명변수 (독립변수)가 두 개 이상인 회기모형



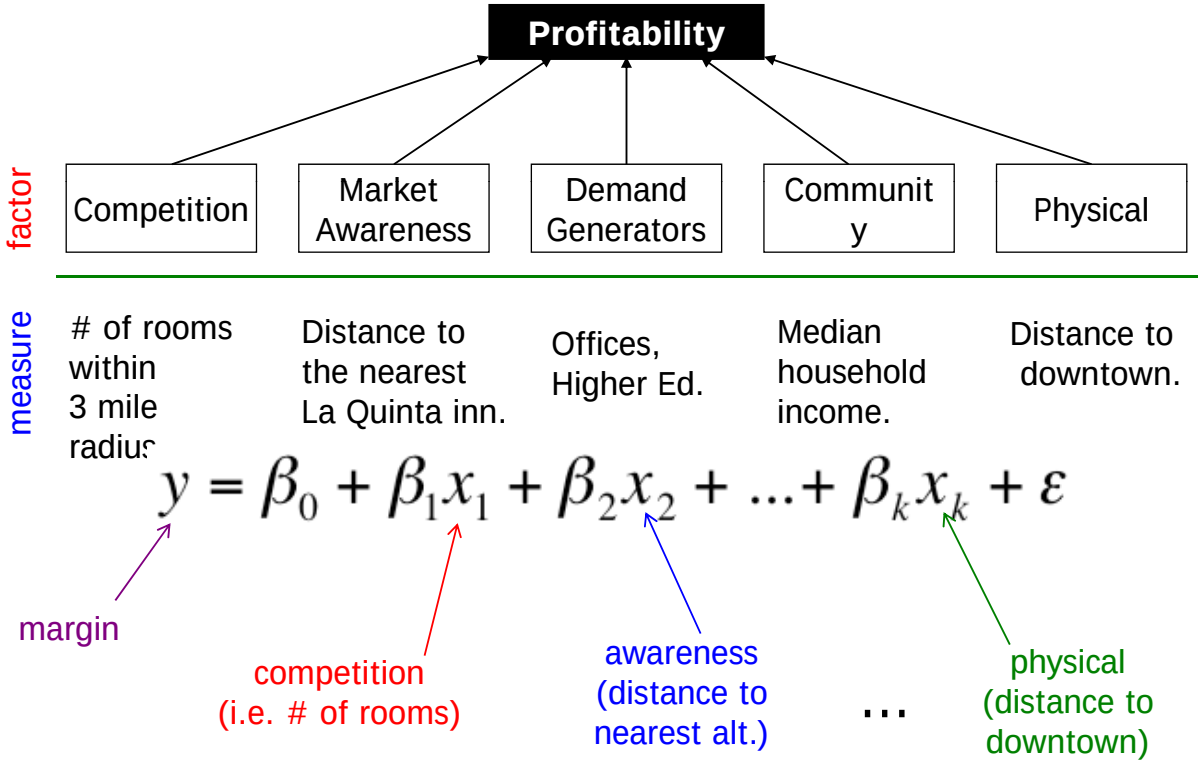
- 회귀분석의 실행순서(단순회귀모형과 동일)
  - (1) 회귀모형의 구성 / 자료수집 / 추정
  - (2) 필요조건의 진단
    - 잔차분석
    - 다중공선성 (multicollinearity)
  - (3) 적합도 평가
    - 회귀계수의 유의성,
    - 결정계수
    - F검정
  - (4) 모형의 해석과 예측

Example

- 예제 16-1; Xm-16-01
  - 신규 설립 모텔의 입지를 선택하기 위하여, 가장 수익성 높은 입지 조건을 조사
  - 종속변수: 모텔의 운영 수익성
  - 설명변수
    - (1) 3마일 이내의 총 객실 수
    - (2) 가장 가까운 경쟁 모텔까지의 거리
    - (3) 주변 지역의 사무실 공간 크기
    - (4) 주변 대학의 등록학생 수
    - (5) 주변 지역사회의 가구 소득 평균
    - (6) 다운타운 중심부까지의 거리
  - 총100개의 모텔을 대상으로 조사

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_6 X_6 + \varepsilon$$

Example



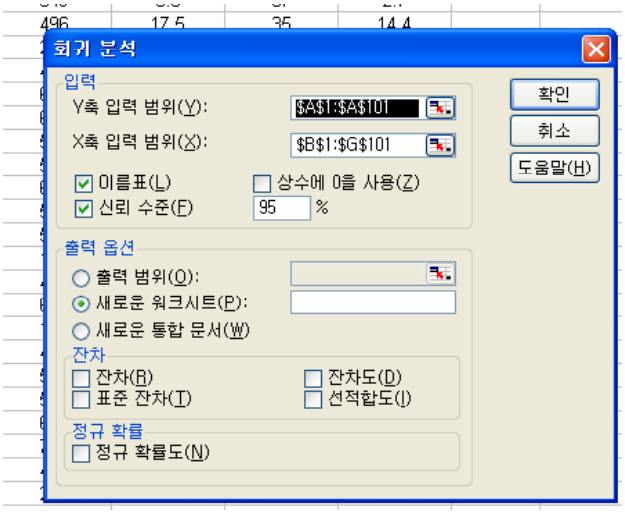
Example

- 추정 결과
  - 도구 – 데이터분석 – 회귀분석
  - **Y축 입력범위 A1:A101**
  - **X축 입력범위 B1:G101**

A	B	C	D	E	F	G
Margin	Number	Nearest	Office Space	Enrollment	Income	Distance
55.5	3203	4.2	549	8.0	37	2.7
33.8	2810	2.8	496	17.5	35	14.4
49.0	2890	2.4	254	20.0	35	2.6
31.9	3422	3.3	434	15.5	38	12.1
57.4	2687	0.9	678	15.5	42	6.9
49.0	3759	2.9	635	19.0	33	10.8
46.0	2341	2.3	580	23.0	29	7.4
50.2	3021	1.7	572	8.5	41	5.5
46.0	2655	1.1	666	22.0	34	8.1
45.5	2691	3.2	519	13.5	46	5.7
44.2	3471	2.2	523	12.0	39	5.4
29.8	3567	2.5	140	13.5	32	9.1
38.4	3264	2.7	404	22.5	29	10.4

Example

- 출력결과



회귀분석 통계량	
다중 상관계	0.724611
결정계수	0.525062
조정된 결정	0.49442
표준 오차	5.512084
관측수	100

분산 분석

	자유도	제곱합	제곱 평균	F 비	유의한 F
회귀	6	3123.832	520.6387	17.13581	3.03E-13
잔차	93	2825.626	30.38307		
계	99	5949.458			

	계수	표준 오차	t 통계량	P-값	하위 95%	상위 95%	하위 95.0%	상위 95.0%
Y 절편	38.13858	6.992948	5.453862	4.04E-07	24.25197	52.02518	24.25197	52.02518
Number	-0.007618	0.001255	-6.068708	2.77E-08	-0.010111	-0.005125	-0.010111	-0.005125
Nearest	1.646237	0.632837	2.601361	0.010803	0.389548	2.902926	0.389548	2.902926
Office Spac	0.019766	0.00341	5.795594	9.24E-08	0.012993	0.026538	0.012993	0.026538
Enrollment	0.211783	0.133428	1.587246	0.115851	-0.053178	0.476744	-0.053178	0.476744
Income	0.413122	0.139552	2.960337	0.003899	0.135999	0.690246	0.135999	0.690246
Distance	-0.225258	0.178709	-1.260475	0.210651	-0.580139	0.129622	-0.580139	0.129622

Goodness of Fit

- 모형의 평가: 과연 모형이 올바른 것인지 어떻게 확인할 것인가?
  - 표준 오차 (Standard error of estimate)
  - 결정계수 (Coefficient of determination)
  - F-검정량 (F-test of the analysis of variance)
- 표준오차

$$s_\varepsilon = \sqrt{\frac{SSE}{n-k-1}}$$

Goodness of Fit

- 결정계수

$$R^2 = 1 - \frac{SSE}{\sum (y_i - \bar{y})^2}$$

$$Adjusted\ \ R^2 = 1 - \frac{SSE/(n-k-1)}{\sum (y_i - \bar{y})^2/(n-1)}$$

Goodness of Fit

$$F > F_{\alpha,k,n-k-1}$$

$$F_{\alpha,k,n-k-1} = F_{.05,6,100-6-1} \approx 2.17 \quad = 0$$

H_1 : 적어도 하나의 $\beta_i \neq 0$

$\Rightarrow$ 검정 통계량 $F = \dfrac{SSR/k}{SSE/(n-k-1)}$

분산 분석

	자유도	제곱합	제곱 평균	F 비	유의한 F
회귀	6	3123.832	520.6387	17.13581	3.03E-13
잔차	93	2825.626	30.38307		
계	99	5949.458			

- 큰 값의 F는 대부분의 y의 변화량이 회기분석 식에 의해 설명되고 있음을 말하며, 즉, 모형이 유의하다는 결과를 의미한다.

Goodness of Fit

- 모형의 유의성

$$F > F_{\alpha,k,n-k-1}$$

$$F_{\alpha,k,n-k-1} = F_{.05,6,100-6-1} \approx 2.17$$

F	Significance F
17.14	0.0000

  - **F = 17.14 이고 our $F_{Critical}$ = 2.17, 따라서 H0를 reject 할 수 있다.**

Adjusted R^2

- 정의
  - 조정결정계수
  - 설명변수의 수가 많으면, 실제의 설명력보다 결정계수의 값이 과대평가될 위험이 존재
    - 이를 조정하기 위해 조정결정 계수 도입
  - 다중 회귀모형에서는 조정결정계수에 의해서 판단

회귀분석 통계량	
다중 상관계수	0.724611
결정계수	0.525062
조정된 결정계수	0.49442
표준 오차	5.512084
관측수	100

Multiple Linear Regression

SSE	S_ε	R^2	F	Assessment of Model
0	**0**	**1**		**Perfect**
small	**small**	**close to 1**	**large**	**Good**
large	**large**	**close to 0**	**small**	**Poor**
$\sum(y_i-\bar{y})^2$	$\sqrt{\frac{\sum(y_i-\bar{y})^2}{n-k-1}}$	**0**	**0**	**Useless**

Test on the Regression Coefficients

- 회귀계수의 유의성 검정
  - 단순 회귀 모형에서처럼 t-검정

	계수	표준 오차	t 통계량	P-값	하위 95%	상위 95%	하위 95.0%	상위 95.0%
Y 절편	38.13858	6.992948	5.453862	4.04E-07	24.25197	52.02518	24.25197	52.02518
Number	-0.007618	0.001255	-6.068708	2.77E-08	-0.010111	-0.005125	-0.010111	-0.005125
Nearest	1.646237	0.632837	2.601361	0.010803	0.389548	2.902926	0.389548	2.902926
Office Space	0.019766	0.00341	5.795594	9.24E-08	0.012993	0.026538	0.012993	0.026538
Enrollment	0.211783	0.133428	1.587246	0.115851	-0.053178	0.476744	-0.053178	0.476744
Income	0.413122	0.139552	2.960337	0.003899	0.135999	0.690246	0.135999	0.690246
Distance	-0.225258	0.178709	-1.260475	0.210651	-0.580139	0.129622	-0.580139	0.129622

- F-검정과 t-검정
  - 단순 회귀모형에서 둘은 동일!
  - 다중 회귀모형에서는 전체적인 회귀 모형의 적합도를 평가하기 위해서는 주로 F-검정을 활용

Residual Analysis – Time Series

- 잔차분석
  - 정규성
  - 이분산성
  - 자기상관
- **Durbin-Watson (DW) test**

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

17.14

Example

- 예제 16-3; Xm 16-03
  - 스키 리프트 매출
  - 종속변수: 스키 리프트 권 판매량
  - 설명변수: 총 강설량 / 평균온도
- DW
  - D=0.5931
  - dL=1.10, dU=1.54
  - 자기상관 존재!
- 해결책
  - 양의 자기상관
  - 1,2,3,…,20으로 구성된 새로운 시간 변수 도입
  - DW=?

Multicollinearity

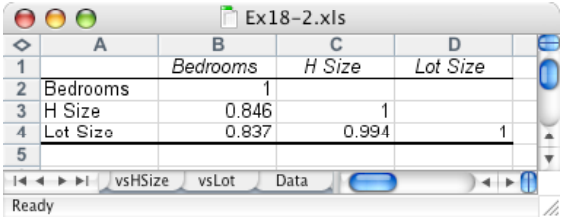
- 정의
  - 다중공선성
  - 다중 회귀모형에서 설명변수들 간에 상관관계가 존재하는 경우
- 부작용
  - (1) 추정된 회귀계수가 실제의 모수 값과 큰 괴리를 나타낼 가능성
    - 회귀계수의 표준오차가 아주 커짐
  - (2) 회귀계수의 유의성 검정에서 대부분 기각하지 못할 가능성
    - 회귀계수의 표준오차가 커지므로 t-값이 작아짐
  - (3) F-검정에서 영향을 미치지 않음

Multicollinearity

- 다중공선성의 진단
  - 설명변수들간의 상관분석을 통한 임의적 진단
  - F-통계량은 큰데 t통계량은 작은 경우 다중공선성의 가능성이 높음
- 다중공선성의 해결
  - 단계별 회기 분석
  - 상관관계가 높은 설명변수의 초기 배제

Example

- 예제 16-2; Xm 16-02
  - 주택판매가격의 예측
  - 종속변수: 주택판매가격
  - 설명변수: 주택크기, 침실의 수, 토지의 크기
- 다중공선성의 진단
  - (1) 상관분석



  - (2) F-검정과 t-검정의 상반된 결과

Exercise

- 연습문제 16-10 (Xr 16-10)
  - 생명보험 고객들의 예상수명
  - 고객 중 최근 사망자 100명 대상으로 조사
  - 종속변수: 고객 수명
  - 설명변수: 부/모의 수명, 조부/모의 수명, 흡연유무

Regression Procedure

- 회기분석 실행 과정
  - **(1) 회기모형 설정 및 추정**
  - **(2) 필요조건 진단**
    - 잔차분석 / 다중공선성
  - **(3) 적합도 평가**
    - 회귀계수의 유의성 / 결정계수 / F-검정
  - **(4) 결과 해석**