

Chapter 1 회귀분석 개념

1.1 기원

회귀(regress)의 사전적 의미는 “go back to an earlier and worse condition” (옛날 상태로 돌아감)을 의미한다. 이런 용어를 사용하게 된 것은 영국의 유전학자 Francis Galton(1822-1911)의 연구에 기인한다. Galton은 (처음에는 sweat pea) 부모의 키와 자녀의 키 사이 관계를 연구하면서 928명의 성인 자녀 키(여자는 키에 1.08배)와 부모 키(아버지 키와 어머니 키의 평균)를 조사하여 다음 표를 얻었다. 이 표를 관찰한 결과 키는 무한정 커지거나 무한정 작아지는 것이 아니라 전체 키 평균으로 돌아가려는 경향이 있다는 것을 발견하였다. 그리하여 그가 제안한 분석 방법의 이름을 “회귀”분석이라 명명하였다.

성인 자녀	부모 키											Totals
	<64.0	64.5	65.5	66.5	67.5	68.5	69.5	70.5	71.5	72.5	>73.0	
>73.7	—	—	—	—	—	—	5	3	2	4	—	14
73.2	—	—	—	—	—	3	4	3	2	2	3	17
72.2	—	—	1	—	4	4	11	4	9	7	1	41
71.2	—	—	2	—	11	18	20	7	4	2	—	64
70.2	—	—	5	4	19	21	25	14	10	1	—	99
69.2	1	2	7	13	38	48	33	18	5	2	—	167
68.2	1	—	7	14	28	34	20	12	3	1	—	120
67.2	2	5	11	17	38	31	27	3	4	—	—	138
66.2	2	5	11	17	36	25	17	1	3	—	—	117
65.2	1	1	4	2	15	16	4	1	1	—	—	48
64.2	4	4	5	5	14	11	16	—	—	—	—	59
63.2	2	4	9	3	5	7	1	1	—	—	—	32
62.2	—	1	—	3	3	—	—	—	—	—	—	7
<61.7	1	1	1	—	—	1	—	1	—	—	—	5
Totals	14	23	66	78	211	219	183	68	43	19	4	928

Galton은 작성한 표를 통하여 부모 키와 자녀의 키 간에는 직선 관계가 있음을 발견하였고 또한 자녀의 키는 평균 키를 중심으로 회귀하려는 경향이 있음을 언급하였다. Galton은 경험적 연구를 통하여 회귀분석 개념을 도출하였다면 Karl Pearson(1903)은 회귀분석 모형과 수학적 전개를 정립화 하였다. Pearson은 1078명의 부자 키를 조사하여 다음 선형 함수 관계를 도출하였다.

$$Y(\text{아버지키}) = 33.73 + 0.516X(\text{아들키})$$

1.2 통계 분석

통계분석은 연구 목적에 의해 수집된 데이터를 분석하여 결론이나 정보를 얻는 일련의 과정이다. 연구 목적이 설정되면 그에 맞는 ①통계적 가설(statistical hypothesis)이나 모형(model)을 설정하고 ②관련 데이터 수집하여 정리하고 분석하여 ③가설 혹은 모형의 유의성(significance)을 검정한다. 이를 **Confirmatory(확증적)** 데이터 분석이라 한다. 대부분의 고전적 데이터 분석 방법은 이에 속하는데 회귀분석도 마찬가지이다.

반면 탐색적 데이터 분석은 수집된 데이터로부터 정보나 결론을 얻는 통계적 분석 방법이다. Tukey의 EDA, 최근의 Data Mining 기법이 탐색적 데이터분석에 속한다.

1.2.1 변수와 데이터

데이터 수집하기 전에 데이터 분석 목적을 설정한다. 데이터를 통하여 어떤 정보를 얻을 것인지 명확하게 설정해야 한다. 이것으로 적절한 표본 추출방법, 변수 설정, 함수 관계 설정, 데이터 수집 방법을 선택이 가능하다. 또한 데이터 역사(data trail, data history)를 수집 과정에(누가, 언제, 어디서, 누구를 대상으로, 어떻게, 수집 과정의 특이한 상황) 대한 상세한 기록을 데이터 역사라 하는데 이는 자료 분석 결과를 이해하고 해석하는데 필수적이다.

분석의 대상이 되는 데이터는 행렬의 형태로 입력되는데 열은 변수(variable), 행은 개체(subject)로 구성되어 있다. 행의 각 원소(셀)을 관측치(observation)라 한다. 변수는 관심의 대상이 되는 개체의 특성(항목)을 의미한다. 변수의 종류에 따라 통계 분석 방법이 결정된다. 다음은 변수의 종류를 정리한 것이다.

■수리 통계

이산형(discrete): 측정 결과를 셀 수 있는 경우이다. 성별, 직업, 교통량, 나이 등이 여기에 해당한다.

연속형(continuous): 측정 결과가 무한이(infinite) 많은 변수를 연속형 변수라 한다. 즉 변수의 범위(range) 중 어떤 구간을 설정하더라도 측정치가 발생할 수 있는 경우로 키, 몸무게, IQ, 소득 등이 여기에 해당된다.

■자료 분석

측정형 변수(metric, measurable, quantitative): 실험 개체의 측정 가능한 특성을 측정할 변수로 키, 몸무게, 평점, IQ, 교통량, 사망자 수가 그 예이다. 연속형 변수는 모두 측정형 변수이고 이산형 변수 중 측정형 변수가 있을 수 있다. 예) 교통 사고 건수, 나이(년)

분류형(범주형) 변수(non-metric, classified, categorical, qualitative): 개체를 분류하기 위해 측정된 변수를 의미하며 성별, 결혼여부 등이 그 예이다.

(1)명목형(nominal): 개체를 분류만 한다. 성별, 결혼여부, 학력

(2)순서형(ordinal): 순서를 가진다. 성적(A, B, ..) 소득수준(상, 중, 하), 리커트 척도(5점)

■시간

데이터가 시간적 순서를 가지면 이를 시계열 자료(time series)라 하고 그렇지 않은 경우를 횡단면 자료(cross-section: 일정 시간에 한꺼번에 조사)라 한다. 경제 지표(환율, 수출량)나 기업의 연차별 자료(연도별 매출액), 주가 등이 시계열 자료에 해당된다. 시계열 다변량 자료에 대한 분석 방법으로는 시계열 자료 분석, 계량 경제(econometrics) 방법(시계열 데이터에 대한 회귀분석) 등이 있다.

■인과 관계

통계 모형의 인과 관계(casual relationship)에서 원인이 되는(영향을 주는) 변수를 설명변수(exploratory variable) 혹은 독립변수(independent)라 하고 결과나 영향을 받는 변수를 종속변수(dependent) 혹은 반응변수(response)라 한다. 종속 변수는 y , 설명 변수는 x 로 표시한다. 분산 분석에서 설명 변수는 처리 효과, 요인으로 불리어진다. 다음은 일반적 통계 모형이다.

$$Y = f(X_1, X_2, \dots, X_p) + e$$

인과 관계는 이론적, 경험적 타당성에 근거하여 데이터 수집 전에 설정되는 것이지 분석 후 설정되는 것은 아니다. 예를 들어 보자. 통계학과 학생 40명의 수능성적과 용돈을 조사하였다. 수능성적을 y , 용돈을 x 로 하여 회귀분석을 실시한 결과 유의한(significant)가 있다고 해서 용돈이 수능성적에 영향을 미친다고 할 수 있는가?

1.2.2 통계소프트웨어

통계 소프트웨어(statistical software)는 수집된 데이터로부터 통계 요약치, 통계량(statistic)을 얻거나 가설검정 및 모형 유의성 검정을 위한 계산 값(확률, 검정통계량, 유의확률) 등을 얻는데 도움을 주는 컴퓨터 소프트웨어이다. 프로그램이 단일 작업을 위하여 작성된 것이라면 소프트웨어는 다수 프로그램들이 하나의 목적을 위해 결합된 형태이다. 통계 소프트웨어의 발달은 데이터 정리 및 분석을 위한 시간 절약 및 계산의 정확성 제고는 물론 통계 비전문가라도 손쉽게 통계 수치를 얻을 수 있게 하였다.

통계 소프트웨어의 종류는 다양하다. 사회과학 분야에서 주로 사용되는 SPSS(Statistical

Package for Social Science), 경영 과학이나 QC 분야의 Minitab, 통계 그래픽에 강한 SYSSTAT, STATGRAPHICS, 시뮬레이션과 수리적 계산에 편리한 S-plus, 경제 데이터 분석의 RATS, EVIEW 등이 있다. 또한 스프레드 시트용 소프트웨어 EXCEL에도 기초적인 데이터 분석 기능이(add-on 기능을 이용하면 엑셀이 기본적으로 제공하는 분석 이외에 보다 다양한 통계분석이 가능) 포함되어 있다. 이처럼 각 통계소프트웨어는 분석 분야에 대한 특성화, 사용 편리화, 작업의 메뉴화를 통하여 특정 분야의 우위를 점하고 있다.

SAS *The Power to Know*[®]의 역사는 미국 North Carolina주 Raleigh에 있는 NCSU(North Carolina State University) 통계학과 대학원 과정 학생들이 주축이 되어 Statistical Analysis System을 완성하던 1966년으로 거슬러 올라간다. 1972년 SAS72가 각 대학에 Shareware 버전으로 제공되어 사용되다가, 1976년 Cary (NCSU에서 15분 거리의 도시)에 SAS Institute를 설립하면서 SAS 제품을 판매되기 시작했다. 초기에는 데이터를 검색하고 통계 분석 및 해석을 위한 소프트웨어였으나 제품이 개발되면서 통합 응용패키지(SAS 약어 바꿈: Strategic Application System)로 발전하였다. 현재 SAS는 전세계 118개국 (미국, 57개국 지사), 40,000 업체(기업, 정부, 연구소, 학교)에서 사용되고 있으며, Fortune 500 기업 중 90%가 SAS를 사용하고 있다. 한국은 version 8, 미국은 version 9을 현재 사용하고 있다.

SPSS Inc.는 1968년 원시 자료(raw data)로부터 기업의 의사결정에 이용되는 정보를 얻기 위한 통계 분석을 위해 개발되어 현재는 SPSS 버전 12가 출시되어 사용되고 있다. 이전에는 영구 라이선스 개념이었으나 얼마 전부터 SAS처럼 라이선스를 일년에 한 번씩 갱신하도록 하는 제품을 출시하고 있다. 라이선스 기간이 끝나도 컴퓨터 날짜를 고치면 사용할 수 있다. SPSS 데이터 확장자는 *.sav이고 출력 결과 확장자는 *.spo이다. 데이터 저장은 필요하나 출력 결과 저장은 큰 의미가 없다.

통계소프트웨어의 발달과 사용의 편리성은 때로 통계 오용과 남용을 불러 일으킨다. 사용자가 데이터를 입력하면 통계 소프트웨어는 그저 하나의 숫자로 인식하게 된다. 그러므로 사용자의 작업 명령에 오류가 없으면 데이터에 적절하지 않은 자료분석 방법이라도 결과를 출력한다. 예를 들어 보자. 평균 평점, 토익 점수가 취업 여부(성공=1, 실패=0)에 영향을 미치는지 알아보기 위하여 데이터를 수집하였다. 적절한 분석 방법이 로지스틱 회귀분석인데도 불구하고 회귀분석을 실시해 보자. 오류 메시지 없이 회귀분석 결과가 출력된다. 이처럼 통계 소프트웨어는 분석 방법의 적절성을 판단할 능력은 없다. 통계 소프트웨어를 이용하여 데이터로부터 정확한 정보를 얻는 것은 이용자의 몫이다.

1.2.3 분석 방법

■일변량 분석

일변량 분석에서 관심의 대상은 모수(parameter θ)에 있으므로 (x_1, x_2, \dots, x_n) 으로부터 숫자

요약인 통계량(statistics: 예, 표본평균, 표본분산, 표본비율 등)을 구하고 그것을 이용하여 모수를 추정하거나 모수에 대한 가설(이를 통계적 가설이라 한다)을 검정하게 된다. 모수 추정에 사용되는 통계량을 추정량(estimator)라 하고 가설검정에 사용되는 통계량을 검정통계량(test statistic)이라 한다.

모수를 추정하는 방법은 점추정(point estimation)과 구간추정(interval estimation)이 있다. 구간추정과 가설 검정을 위해서는 추정치와 검정통계량의 분포에 대한 정보가 필요하다. why? 분포를 알아야 확률을 구할 수 있기 때문이다. (95% 신뢰구간, 5% 유의수준, 유의확률) 모집단의 분포를 모르고 소표본인 경우 검정통계량의 분포를 알 수 없으므로 비모수 방법(nonparametric, distribution free)을 사용하게 된다. 일변량 분석 데이터 변수들을 분석하는 방법으로 정리하면 다음과 같다.

	그래프 요약	숫자 요약
측정형	줄기 잎(stem and leaf) 그림 상자 수염(box-whisker) 그림 히스토그램(histogram)	기초통계량: (평균, 표준 편차) 중앙값, 범위, 사분위, IQR
	데이터 분포 형태 (치우침, 봉우리 개수), 이상치 발견	(*모집단 평균을 위한 t-검정 (*비모수 방법(Sign test, Wilcoxon rank sum test) (*모집단 분산에 대한 χ^2 -검정
범주형	바(bar) 차트, 파이(pie) 차트	비율(proportion, ratio)
	각 수준에 빈도나 상대 빈도(비율)를 그래프로 나타낸다.	모집단 비율에 대한 신뢰구간 및 가설 검정 (*모비율에 대한 z-검정(대표본) (*유의확률(p-value) 이용(소표본)
(*모비율에 대한 z-검정(대표본)		
	각 수준에 빈도나 상대 빈도(비율)를 그래프로 나타낸다.	모집단 비율에 대한 신뢰구간 및 가설 검정 (*모비율에 대한 z-검정(대표본) (*유의확률(p-value) 이용(소표본)

두 모집단 평균(t-검정) 및 비율(z-검정) 차이 비교, 모분산 차이 검정(F-검정), 짝진 표본 평균 차이 검정(t-검정)은 일변량 분석에서 일반적으로 다룬다. 짝진 표본검정은 짝 이룬 데이터의 차이를 구한 후 모집단 평균이 0인지를 검정하면 된다. 다음은 두 모집단 비교를 위한 분석 방법을 정리한 것이다.

	그래프 요약	숫자 요약
측정형	상자 수염(box-whisker) 그림	기초통계량: (평균, 표준 편차) 중앙값, 범위, 사분위, IQR

	데이터 분포 형태 이상치 발견, 분산의 차이 정도	(*)모집단 평균 차이 t-검정 (*)모집단 분산 차이 F-검정
범주형	바(bar) 차트	비율(proportion, ratio)
집단별 상대 빈도를 그래프		
(*)모비율 차이에 대한 z-검정(대표본)		
(*)교차 분석(소 표본)		

단일 집단 모평균 추론 $H_0: \mu = \mu_0$

☑검정통계량: $T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t(n-1) \text{ or } N(0,1)$

☑신뢰구간: $\bar{x} \pm t(n-1; 1-\alpha/2) \frac{s}{\sqrt{n}}$

단일 집단 모비율 추론 $H_0: p = p_0$

☑검정통계량: $T = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \sim N(0,1)$ (대표본), 소표본은 유의확률 개념을 이용한다.

☑신뢰구간: $\hat{p} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$

두 집단 모평균 추론 $H_0: \mu_1 = \mu_2$

☑검정통계량: $T = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2) = 0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n-2) \text{ or } N(0,1)$

통합분산 $s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$

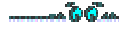
☑신뢰구간: $\bar{x}_1 - \bar{x}_2 \pm t(n-2; 1-\alpha/2) s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

두 집단 모비율 추론 $H_0: p_1 = p_2$

☑검정통계량: $T = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2) = 0}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}} \sim N(0,1)$ (대표본)

소표본은 교차분석의 EXACT test 이용

☑신뢰구간: $T = \hat{p}_1 - \hat{p}_2 \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$



▣다변량 분석

다변량 분석은 세 개 이상(이변량 분석은 변수 2개 관계) 변수간의 1)인과 관계(casual relationship)를 규명, 분석하거나 (회귀 분석: regression, (다변량) 분산 분석: (Multivariate) ANOVA) 2)상관 관계(correlation)를 이용하여 변수 축약(reduction)이나 개체 분류(classification)에 관련된 분석 방법이다. 다음은 인과 관계를 분석하는 다변량 분석을 변수 형태별로 정리하면 다음과 같다.

설명 변수 \ 종속변수	측정형	분류형
측정형 (측정형+분류형)	회귀분석 (*상관분석: 상관 관계만	로지스틱 회귀분석(이진 종속변수) Logit 회귀분석(순서형 종속변수)
분류형	분산 분석	교차 분석(χ^2 -검정) 범주형 자료분석: Log-normal 모형

일반적으로 협의의 다변량 분석이란 변수의 상관 관계를 이용하여 대용량 데이터의 변수를 축약하거나 개체들을 몇 개의 그룹으로 나누는데 사용되는 분석 방법을 일컫는다. 일반적으로 다변량 분석은 변수를 축약(주성분분석)하거나 그룹화(요인분석) 하는 변수유도기법과 개체의 집단을 판별(판별분석)하거나 분류하는(군집분석, 다차원척도법, 대응분석) 개체유도기법으로 나뉜다.

1.2.4 일변량 분석 예제



EXAMPLE 1-1

일변량 분석: 모집단 평균 검정

EXAMPLE1-3 데이터 이용: 물에 의해 식히면 철강의 강도가 145라고 한다. 오일에 의해 식히는 경우 철강의 강도가 145보다 크다고 할 수 있나? 유의수준 5%



EXAMPLE 1-2

일변량 분석: 두 모집단 평균 차이 검정(독립)

철강 생산 시 식히는 과정에서 소금 물을 사용하는 방법과 오일을 사용하는 방법 중 어느 것이 강도를 높이는지 알아보기 위하여 다음과 같이 측정 자료를 얻었다. 강도는 정규 분포를 따른다고 하자.

소금물: 145 150 153 148 141 152 146 154 139 148

오일: 152 150 147 155 140 146 158 152 151 143

**EXAMPLE 1-3**

일변량 분석: 두 모집단 평균 차이 검정(짜진)

두 시험지의 난이도 차이가 있는지 알아보기 위하여 12명의 학생을 임의 추출하여 실험을 본 결과이다. 학생 한 명으로부터 시험 점수가 각각 나왔으므로 짜진 표본이다. 데이터 입력할 때 주의해야 할 것은 짜진 데이터를 2개의 열에 입력하면 된다. 첫 번째 행의 관측치는 학생1의 시험1 성적과 시험2 성적을 입력한 것이다.

	시험1	시험2
1	33	41
2	37	41
3	30	46
4	42	46
5	48	53
6	43	39
7	45	49
8	35	45
9	57	39
10	37	45
11	40	49
12	49	54

**EXAMPLE 1-4**

일변량 분석: 모비율 차이 검정

하루에 생산된 제품의 불량률이 10%보다 크면 그날의 모든 제품을 폐기한다고 한다. 오늘 생산된 제품 100개를 조사하였더니 불량률의 개수가 15개였다. 그럼 오늘 제품을 모두 폐기해야 하는가? 유의수준은 0.05로 하시오.

표본의 크기를 15개로 하였더니 불량 제품 개수가 2개였다. 제품 모드를 폐기해야 하는가? 유의수준은 0.05로 하시오.

**EXAMPLE 1-5**

일변량 분석: 두 모집단 비율 차이 검정

어떤 사람이 CEO들이 일반 사람에 비해 왼손잡이가 많다고 주장하였다. 일반인 100명 중 85명, CEO의 300명 중 270명이 왼손잡이였다. 유의수준 0.05에서 그의 주장이 옳은지 판단하시오.

**HOMEWORK #0**

DUE 3월7일(월)

각자 컴퓨터에 SAS(8.2/라이선스 홈페이지)와 SPSS(12/조교실)를 설치하시오.

1.3 회귀분석이란?

1.3.1 일반 개념

회귀분석은 이론이나 경험적 근거에 의해 설정된 변수들간의 (선형) 함수관계가 유의한 지 알아보는 통계분석 방법이다. 회귀분석은 다음 물음에 답할 수 있다.

- ① 변수들간의 함수 관계가 존재하는지?(일반적으로 다루기 쉽고 해석이 용이한 선형 함수를 가정한다) 함수 관계($Y = f(X_1, X_2, \dots, X_p)$)에서 Y변수를 종속변수(dependent variable)라 하고 X변수를 설명(exploratory)변수 혹은 독립(independent)변수라 한다.

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + e$$

- ② 설정된 설명변수의 영향(설명)이 유의(significant)한지? ($H_0 : b_k = 0$) 유의한 설명변수 중 종속변수의 영향력이 가장 큰 것은?(표준화 회귀계수 B_k 의 크기) 종속변수에 대한 설명변수의 유의성과 영향력은 회귀계수를 이용한다.
- ③ 최종 회귀모형을 이용하여 주어진 설명변수에 있어서 종속변수 예측치를 얻는다.

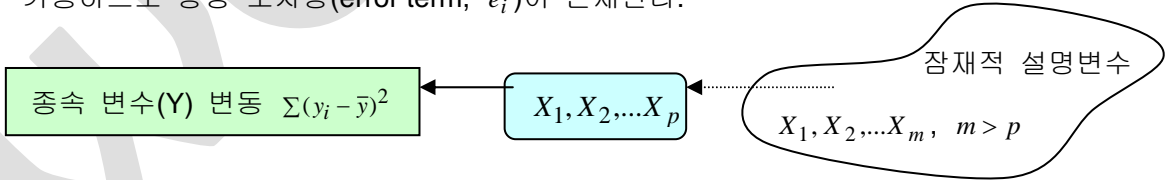
$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_p X_p$$

회귀분석 과정을 정리하면 다음과 같다.

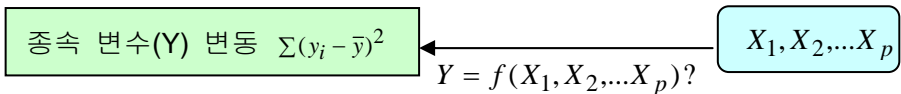
- ① 종속 변수의 무엇을 설명할 것인가? 종속 변수가 가진 정보를 어떻게 표현 할 것인가? 통계학에서는 종속 변수가 가진 정보를 변수의 변동(분산)으로 생각했다.

종속 변수(Y) 변동 $\sum (y_i - \bar{y})^2$

- ② 어떤 설명변수(독립 변수)가 종속 변수에 영향을 미칠까? 변수의 선택은 이론이나 경험에 의해 분석자가 선택하게 된다. 영향을 미치는 모든 설명변수를 고려하는 것은 불가능하므로 항상 오차항(error term, e_i)이 존재한다.



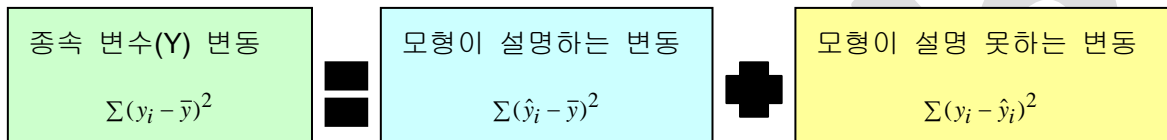
- ③ 설명 변수와 종속 변수는 어떤 함수 관계를 갖는가? 실제 함수 f 의 형태는 알 수 없거나 이론 모형은 복잡하다(nonlinear: 비선형 회귀 모형) 그리하여 함수 관계를 단순화 하여 다루기 쉽고 해석이 용이한 선형 함수를 선택하게 된다.



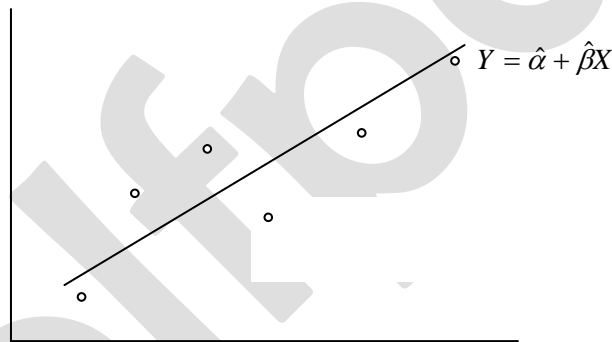
변수들간의 어떤 함수 형태가 존재하는지? (선형 함수: $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi} + e_i$)
영향을 미치는 설명변수는 어떤 것이며 어느 설명변수의 영향력이 가장 큰가? (모형의 유의성 검정($H_0: \beta_i = 0$ for all i): F-검정, 회귀계수에 대한 유의성 검정(t-검정, 표준화 회귀계수)

④데이터 $(y_i, x_{1i}, x_{2i}, \dots, x_{pi})$, $i = 1, 2, \dots, n$ 으로부터 선형 모형의 회귀계수를 추정한다.

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_p X_{pi}$$



수집한 데이터는 우선 변수들간의 함수 관계를 보기 위하여 산점도(scatter plot)를 그린다. 회귀분석의 시작은 산점도를 그리면서 시작된다. 데이터에 가장 적합한 직선을 어떻게 그을 것인가? 가장 많이 사용되는 방법은 OLS(Ordinary Least Square, 최소자승법) 방법이다. 이렇게 얻은 회귀식($Y = \hat{\alpha} + \hat{\beta}X$)이 유의한가를 검정하게 된다.

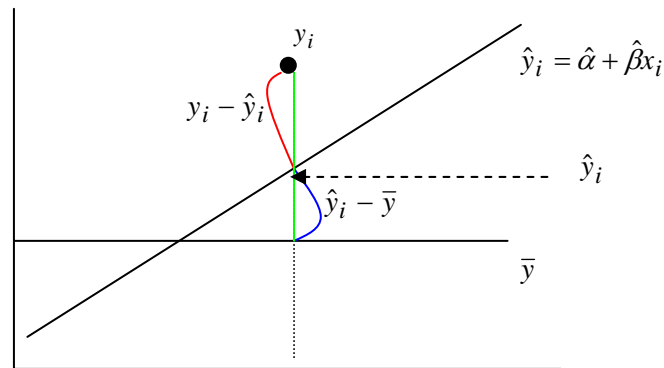


회귀모형의 유의성은 (1)회귀모형에 고려한 설명변수 중 적어도 하나는 유의한지를 알아보는 분산분석(F-검정), (2)개별 설명변수에 대한 유의성은 t-검정을 실시한다.

1.3.2 분산 분석적

회귀 분석을 분산 분석적 측면에서 다루는 것은 단순 회귀에서는 새로운 것이 없으나(단순 회귀 모형에서는 회귀 계수에 대한 t-검정은 분산 분석의 F-검정과 동일하다. $F(1, n-2) = t^2(n-2)$) 보다 복잡한 회귀 모형을 다루는데 도움을 얻을 것이다.

분산분석 접근은 종속변수 Y에 관련된 총변동을 자유도 분할에 근거하여 서로 독립인 두 개의 변동으로 나눈다. 총변동(SSTO, SST, Total Sum of Square)은 종속 변수의 관측치와 평균의 편차(deviation) $(y_i - \bar{y})$ 제곱 합을 의미하며 이는 종속 변수가 가진 정보이다.



총변동은 $SSTO = \sum (y_i - \bar{y})^2 = (\text{간편식}) \sum Y_i^2 - \frac{(\sum Y_i)^2}{n} = \sum Y_i^2 - n\bar{Y}^2$ (초록색 부분)으로 정의된다. 총변동의 자유도 $(n-1)$ 이다.

회귀 모형에서 데이터에 포함된 불확실성(uncertainty)은 적합 회귀선(fitted regression line, 추정 회귀식)으로부터 관측치가 얼마나 벗어나 있나를 의미하며 이것에 대한 측정은 $(y_i - \hat{y}_i)$ 이고 제곱 합을 오차변동(Error Sum of Squares, SSE) 혹은 오차 제곱합(자승합)이라 하며 적합 회귀식에 의해 설명되지 않는 변동에 해당된다. 이 오차 변동을 $(n-p-1)$ 으로 나눈 값을 MSE (Mean Squared Error)라 하면 이는 오차의 분산에 대한 추정치로 사용한다. SSE의 분포는? $\chi^2(n-p-1)$ why? 오차 변동: $SSE = \sum (y_i - \hat{y}_i)^2$ (빨간 부분)


두 변동의 차이를 회귀변동(Regression Sum of Squares, SSR) 혹은 모형 변동(Model SS)이라 하며 적합한 회귀식이 데이터의 관계를 얼마나 잘 설명하는지 나타낸다. SSR의 자유도는 총변동의 자유도-오차변동의 자유도이다. SSR/p 을 MSR (Mean Squared Regression)이라 한다. SSR의 분포는? $\chi^2(p)$ why? 회귀(모형) 변동: $SSE = \sum (\hat{y}_i - \bar{y}_i)^2$ (파란 부분)

그러므로 통계량 $T = \frac{MSR}{MSE}$ 은 $F(p, n-p-1)$ 분포를 따른다. 이 검정통계량을 이용하여 회귀모형에 설정된 설명변수의 유의성(종속변수를 설명하는지)을 검정하게 된다. F-검정의 귀무가설은 “모든 설명변수는 유의하지 않다($H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$)”이다.

각 개별 설명변수의 회귀계수 유의성($H_0: \beta_k = 0$) t-검정을 이용한다.

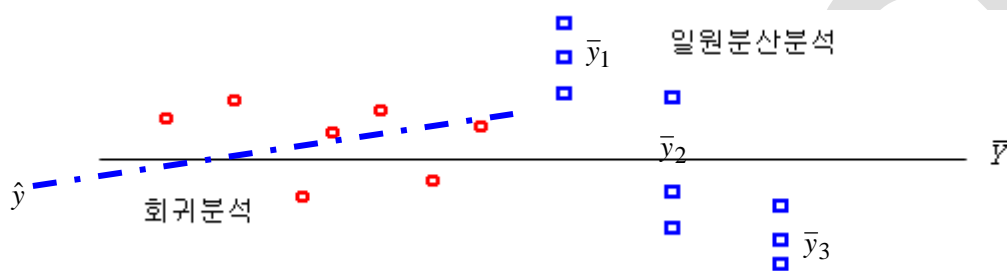
$$T = \frac{\hat{\beta}_k - \beta_k}{s_{\hat{\beta}_k}} \sim t(n-p-1)$$


1.3.3 분산분석과 비교

폐암 사망지수(측정형)에 영향을 미치는 변수로 직업 종류(A/B/C, 분류형, 명목형)와 흡연지수(측정형)를 고려하였다. 이 경우 폐암 사망지수는 종속변수, 직업의 종류와 흡연지수는 독립변수(설명변수)에 해당된다. 설명변수 모두들 고려한 지시변수(indicator)가 있는 회귀분석을 실시하면 되나 여기서는 분산분석과 회귀분석 비교를 위하여 설명변수가 하나인 경우에 한정하겠다. 데이터  SMOKING.XLS (엑셀 데이터)

회귀모형: $y_i(\text{사망지수}) = \alpha + \beta * x_i(\text{흡연지수}) + e_i$

일원분산분석 모형: $y_{ij}(\text{사망지수}) = \mu + \tau_i + e_{ij}, i=1, 2, 3(\text{수준}), j=1, 2, \dots, n_i(\text{반복 수})$




 SAS 엑셀 창에서 데이터 부분을 복사하고(CTRL+C) SAS 확장편집기(Program editor) 창에 붙여 넣기(CTRL+V) 한 후 SAS 데이터 만들기 단계를 조정하면 된다.

```

확장 편집기 - 제목없음1 * PROC REG 실행
data sm;
  input Group $ Smoking Mortality;
  cards;
A 77 84

```

 SPSS 엑셀 창에서 데이터 부분을 복사하고(CTRL+C) SAS 확장편집기(Program editor) 창에 붙여 넣기(CTRL+V) 하면 된다. SPSS도 스프레드시트 데이터 입력 창이므로 이것이 가능하다. 그러나 주의할 것이 하나 있다. 그냥 붙여 넣기를 하면 문자열인 경우에는 결측치가 되므로 창 아래 “변수보기” 폴더에서 문자열이 변수는 유형에서 문자열로 고쳐주자.

	직업	흡연지수	사망지수
1	A	77	84
2	A	137	116
3	A	117	123
4	A	94	128
5	B	116	100

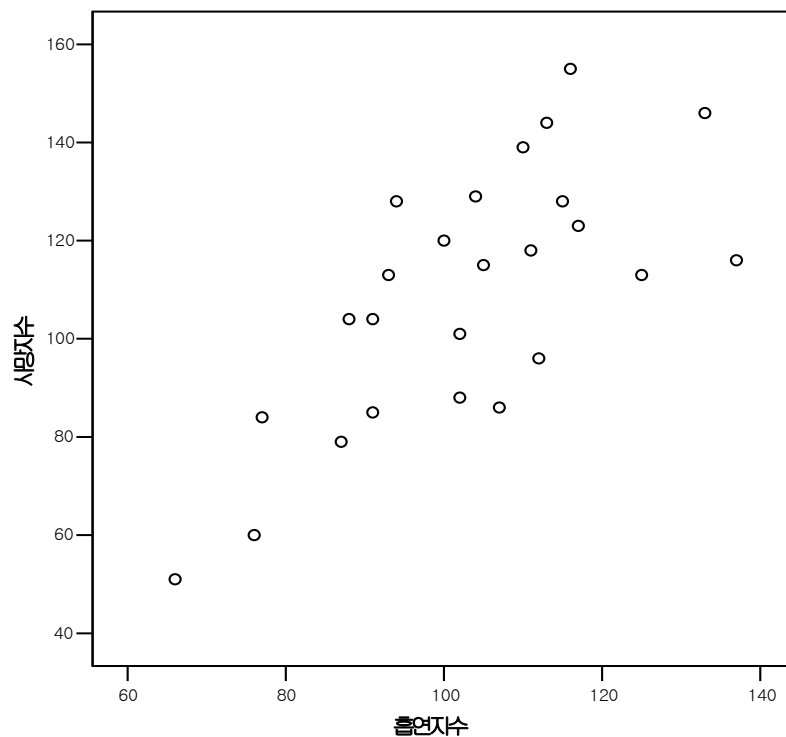
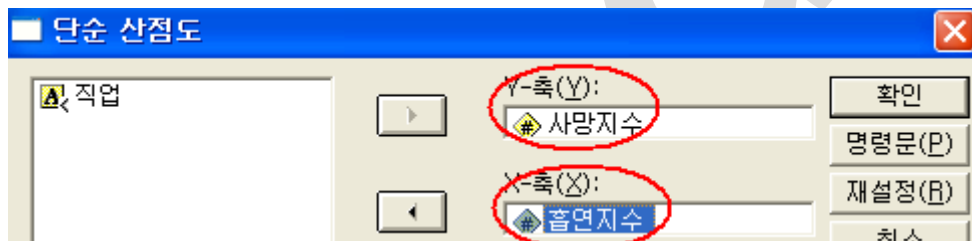
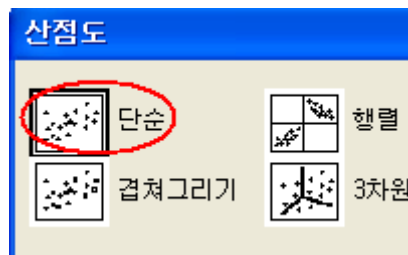
▶ 데이터 보기 / 변수 보기 /

회귀분석의 시작은 산점도(Y축을 종속변수, X축을 설명변수)를 그리면서 시작한다.


SAS SAS는 회귀분석 모형 추정과 산점도를 하나의 프로그램을 그릴 수 있다.

```
proc reg data=sm;
  model Mortality=Smoking;
  plot Mortality*Smoking;
run;
```

SPSS 메뉴에서 **그래프(G) ▶ 산점도(S)...** 선택하고 다음과 같이 메뉴를 설정한다.



산점도를 통하여 흡연지수가 높아지면 사망지수가 높아지는 경향을 보이고 있음을 알 수 있다. 분산이 커지는 문제(이분산성)는 존재하지만... 실제 회귀모형이 통계적으로 유의한지 알아보려면 회귀계수 추정과 검정을 실시해야 한다.

 이전 프로그램 실행하여 산점도와 함께 출력 창에 얻은 결과이다.

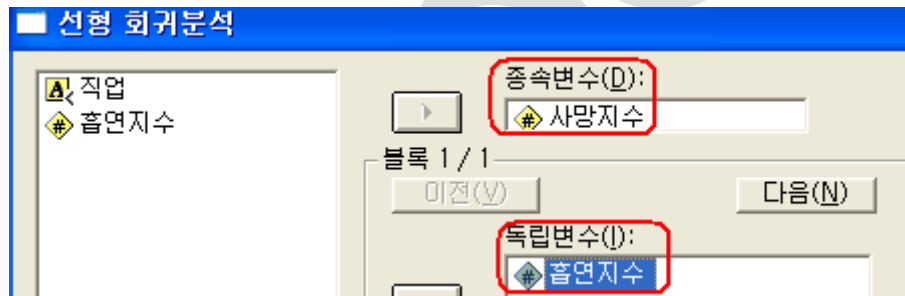
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	8395.74904	8395.74904	24.23	<.0001
Error	23	7970.25096	346.53265		
Corrected Total	24	16366			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-2.88532	23.03372	-0.13	0.9014
Smoking	1	1.08753	0.22095	4.92	<.0001

 메뉴 **분석(A)** ▶ **회귀분석(R)** ▶ **선형(L)...** 선택한 후 아래와 같이 창을 설정하면 된다.



계수^a

모형	비표준화 계수		표준화 계수	t	유의확률
	B	표준오차	베타		
1 (상수)	-2.885	23.034		-0.125	.901
흡연지수	1.088	.221	.716	4.922	.000

a. 종속변수: 사망지수

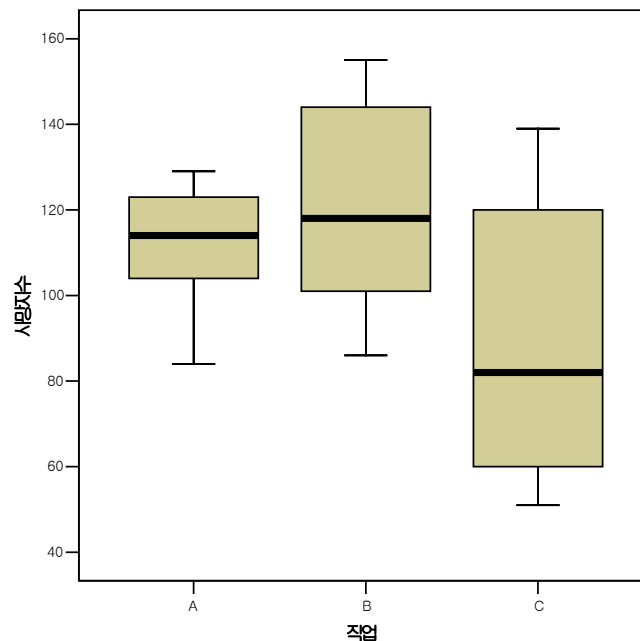
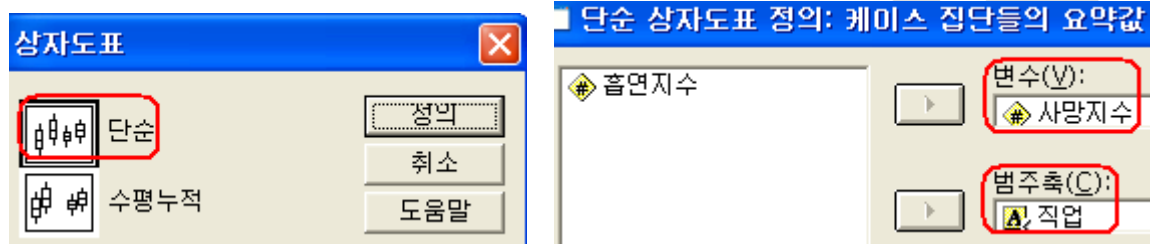
유의확률이 0.001보다 작으므로 흡연지수는 사망지수에 유의한 영향을 미친다. 회귀계수가 양이므로 흡연지수가 높을수록 사망지수가 높아짐을 알 수 있다. 흡연지수가 1단위 증가하면 사망지수는 1.088 증가한다. 그리고 흡연지수가 100인 사람은 사망지수는 105.92이다.

$$\text{사망지수} = -2.885 + 1.088 * \text{흡연지수}$$

설명변수가 하나인 경우에는 회귀계수 유의성 검정의 t-통계량의 제곱이 회귀모형 유의성 검정의 F-검정과 같고 유의확률은 동일하다.

직업 종류가 폐암 사망지수에 영향을 미치는지 보려면 실험 설계를 해야 한다. 각 직업에서 건강 정도가 동일한(실험 전에 집단 간 건강 차이는 없다) 사람을 선택하여 일정 기간이 지난 후 폐암 사망지수를 측정해야 한다. 그러나 여기서는 사후적 관찰을 통하여 직업의 영향을 살펴보고자 한다. 실험 설계에서는 직업이 처리효과(treatment effect) 혹은 요인(factor)이 된다. 처리 효과에 따른 종속변수의 차이를 본다는 것은 처리효과 수준의 종속변수 평균 차이 검정과 동일하다. 그러므로 처리효과(요인) 수준이 2개인 경우에는 독립인 두 모집단 평균 차이 비교와 동일하다.

SPSS 메뉴에서 **그래프(G)** ▶ **상자도표(X)...** 선택하고 아래와 같이 창을 설정한다.



직업 C군의 사망지수가 가장 낮고 A, B 군이 비슷해 보인다. 실제 직업간 차이가 있는지 알아보려면 분산분석(F-검정)을 실시한다. 직업(처리효과) 수준이 3개이므로 두 수준간 쌍체비교(pairwise comparison, 이를 사후분석, post-hoc test이라 한다)를 위하여 다중비교 분석을 실시한다.



```
proc glm data=sm;
  class group;
  model Mortality=group;
  means group/scheffe tukey lines;
run;
```

분산분석 결과 유의확률이 0.06이므로 귀무가설은 채택되어 직업에 따른 사망지수 차이는 없다고 할 수 있다.

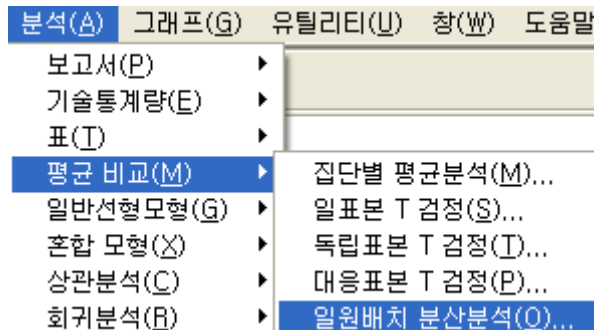
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3668.04444	1834.02222	3.18	0.0613
Error	22	12697.95556	577.17980		
Corrected Total	24	16366.00000			

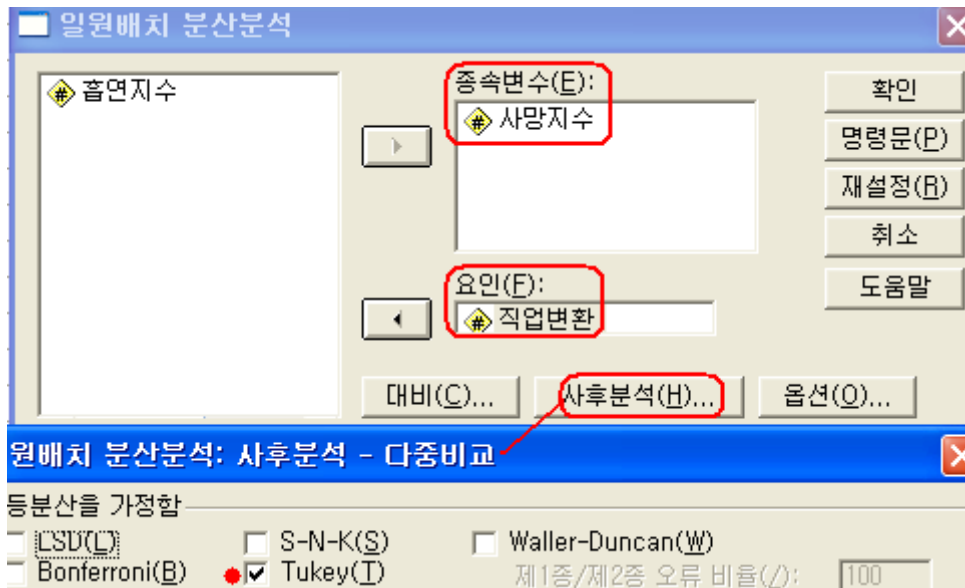
비록 F-검정 결과 차이가 없더라도 사후 검정을 반드시 한다. 직업B의 사망지수는 120.8, 직업A=110.4, 직업C=89이다. 각 수준 간 차이가 통계적으로 유의한가? **LINES** 옵션에 의해 앞에 알파벳이 나타난다. 알파벳이 같은 수준은 차이가 없음을 의미한다. 그러므로 직업 (B, A), (A, C) 차이가 없다. 그러나 (B, C) 차이는 유의하다.

Tukey Grouping	Mean	N	Group
A	120.78	9	B
A	110.40	10	A
B	89.00	6	C



SPSS에서 요인으로 사용되는 변수는 숫자로 입력되어 있어야 한다. 그래서 메뉴에서 **변환(T) ▶ 자동 코딩변경(A)...**을 선택하고 새 이름을 “직업변환”으로 하여 새로운 변수(숫자형)를 만든 후 분산분석 절차를 시행하였다.





분산분석

사망지수

	제 곱합	자유도	평균제 곱	F	유의확률
집단-간	3668,044	2	1834,022	3,178	,061
집단-내	12697,956	22	577,180		
합계	16366,000	24			

종속변수: 사망지수
Tukey HSD

다중 비교

(I) 직업변환	(J) 직업변환	평균차 (I-J)	표준오차	유의 확률	95% 신뢰구간	
					하한값	상한값
A	B	-10,378	11,039	,621	-38,11	17,35
	C	21,400	12,406	,219	-9,77	52,57
B	A	10,378	11,039	,621	-17,35	38,11
	C	31,778	12,662	,050	-,03	63,59
C	A	-21,400	12,406	,219	-52,57	9,77
	B	-31,778	12,662	,050	-63,59	,03

회귀분석의 분산분석 결과와 일원 분산분석의 분산분석 결과를 비교해 보자. 수정 총변동(corrected total variance $SSTO = \sum (y_i - \bar{y})^2$)은 동일하다. Why? 총변동이 두 개의 독립적인 변동으로 나뉘어지는데 오차 변동과 회귀분석은 회귀모형(직선)이 설명하는 모형 변동, 일원 분산분석은 수준간 평균의 차이에 의한 설명인 요인 변동으로 나뉜다.

Source	회귀분석 결과	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model		1	8395,74904	8395,74904	24,23	<.0001
Error		23	7970,25096	346,53265		
Corrected Total		24	16366			



HOMEWORK #1-1

DUE 3월 9일(수)

회귀모형 $SSTO = \sum (y_i - \bar{y})^2$ 과 일원분산분석 $SSTO = \sum \sum (y_{ij} - \bar{y})^2$ 변동 분할 식 유도

$$\text{회귀모형 } \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2,$$

$$\text{일원분산분석 } \sum_{i=1}^k \sum_{j=1}^n (y_{ij} - \bar{y})^2 = \sum_{i=1}^k \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^k \sum_{j=1}^n (\bar{y}_i - \bar{y})^2 \text{ (반복 수 } n \text{으로 동일)}$$



HOMEWORK #1-2

DUE 3월 9일(수)

AD.XLS (엑셀 데이터)

- (1) 광고비(spend)가 상품 인지도(rate)에 영향을 미치는지 회귀분석 하시오.
- (2) 기업 종류(group)에 따라 상품 인지도의 차이가 있는지 일원 분산분석 하시오,

1.3.4 시계열 데이터 회귀분석

시계열(time series) 데이터는 관측치가 시간적 순서를 가지게 된다. 일정 시점에 조사된 데이터는 횡단(cross-sectional) 자료라 한다. ○○전자 주가, △△기업 월별 매출액, 소매물가지수, 실업률, 환율 등이 시계열 자료이다. 시계열 데이터에 대한 회귀분석을 ECONOMETRICS(계량경제)이라 한다. 다음 예제를 통해 간단히 살펴보자.



EXAMPLE 1-6

시계열 데이터 회귀분석

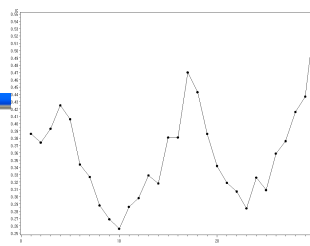
ICECREAM.txt (텍스트 데이터)

아이스크림의 소비량에 미치는 요인으로 가격, 소득, 기온을 생각했다. 29주 동안 주별 데이터를 수집한 결과이다. 회귀분석을 실시해 보자.

시계열 데이터는 산점도보다 종속변수에 대한 시간도표(time plot)가 더 효율적이다.



```
proc gplot data=icecream;
  symbol i=join v=dot;
  plot ic*date;
run;
```



```

proc reg data=icecream;
  model ic=price income temp;
run;

```

분산분석의 F-검정 결과 유의확률이 0.001로 유의하므로 설정한 설명변수 3개 중 적어도 하나는 종속변수에 대한 설명력이 유의하다. 회귀계수 t-검정 결과 가격(PRICE)은 아이스크림 소비량에 영향을 미치지 않는다.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	0.09025	0.03008	22.17	<.0001
Error	26	0.03527	0.00136		
Corrected Total	29	0.12552			

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	0.19732	0.27022	0.73	0.4718
price	1	-1.04441	0.83436	-1.25	0.2218
income	1	0.00331	0.00117	2.82	0.0090
temp	1	0.00346	0.00044555	7.76	<.0001

설명변수 PRICE를 제외하고 회귀 분석한 결과 다음을 얻었다. 물론 아직 검정해야 할 것이 남았지만(오차 자기 상관, 이상치 점검, 잔차분석 등)... 뒤로 미루고 최종 회귀모형은

$$Y_t = -0.01132 + 0.00353 * IN_t + 0.0035TEMP_t$$

소득이 높아질수록 온도가 높을수록 아이스크림 소비량은 높아짐을 알 수 있다.

Variable	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	-0.11320	0.10828	-1.05	0.3051
income	0.00353	0.00117	3.02	0.0055
temp	0.00354	0.00044496	7.96	<.0001



HOMEWORK #2-1

DUE 3월 16일(수)

ICECREAM.txt (텍스트 데이터)

아이스크림 소비량에 소득이 영향을 미칠 것이다. 그러나 주급을 받을 것을 예상하여 소비하지는 않을 것이다. 즉 지난 주 소득이 이번 주 아이스크림 소비량에 영향을 미칠 것이다. 그러므로 설명변수를 지난 주 소득, 가격, 온도를 설명변수로 하여 위의 작업을 실시해 보시오.