통계적 방법론의 이해

성내경 교수 (이화여자대학교)

1. 통계의 기본 개념

통계학은 자료를 수집하고 해석하는 학문이다.

통계학은 확실히 예측할 수 없는 현상에 대하여 자료를 수집하고, 수집된 자료에 담긴 정보를 해석하는 절차이다.

자료수집: 통계조사(survey), 비교실험(comparative experiment)

합리적이고 객관적인 통계적 결론을 내리려면 **랜덤화**(randomization)를 해야한다.

「주어진 자료에 대한 유일하고 적합한 자료 분석법은 없다.」

통계조사의 핵심은 비용과 시간을 절감하면서 조사 대상을 잘 대표할 수 있는 일부의 자료를 선택하여 분석하는데 있다. 조사 대상 전부를 **모집단**(population)이라고 하며, 이 중에서 선택된 일부를 **표본**(sample)이라 한다.

통계학은 몇 그루의 나무를 보고 산 전체를 보려는 학문이다. 부분을 보고 전체를 판단할 때는 반드시 착오를 범할 확률이 존재한다. 그러나, 통계학이 가치있는 이유는 가능한한 오차를 줄이면서 최소의 표본으로 모집단의 성질을 유추할 수 있는 합리적인 도구이기 때문이다.

자료의 해석, 혹은 통계 분석 분야는 두 개의 큰 가지로 나뉘어진다. 하나는 자료의 기술(description)에 관한 **기술통계**(descriptive statistics), 또다른 하나는 자료로부터 모집단의 특성을 추정 및 검증하는 **통계추론**(statistical inferences)이다.

자료기술 기법은 크게 나누어, 막대그래프, 파이그림 등 그래프를 이용하는 기법과 평균, 분산, 중앙값 등 수치화된 기술통계량을 사용하는 기법이 있다. 이 두 기법을 종합한 상자그림(box plot)이라 불리우는 기법이 최근에 폭넓게 사용되고 있다.

통계추론이 필요한 경우는 모집단의 일부만 가용할 때이다. 이때는 모집단을 대표하는 표본을 추출하여 표본이 추출된 모집단에 관한 결론을 끌어낸다.

통계학에서 컴퓨터의 역할은 나날이 증대되고 있다. 수집된 자료의 수가 적으면 손으로 계산할수도 있지만, 태반의 경우 손으로 직접 계산은 거의 불가능하다. 게다가 개인용 컴퓨터

의 보급이 확산되면서 누구나 컴퓨터를 이용하여 손쉽게 자료 처리를 할 수 있게 되었다.

일반적으로 통계 분석을 스스로 수행하려면, 반드시 컴퓨터 사용법에 대하여도 어느 정도의 지식을 갖추어야 한다. 요즈음에는 어느 기관이나 전문적으로 수집된 자료를 통계 처리하는 패키지 프로그램이 최소한 하나는 설치되어 있다. 많이 사용되는 통계 패키지 프로그램으로는 SAS, MINITAB, S-Plus, SPSS 등이 있다.

그러나, 통계 패키지 프로그램의 사용법을 안다는 것이 곧바로 자료분석을 의미하지는 않는다. 패키지 프로그램의 보급이 늘어나면서 한두 페이지의 출력으로 충분한 경우에도 불구하고, 무분별하게 산더미같은 출력을 들고와 자료해석을 요청하는 경우가 종종 있다. 프로그램은 인간이 하라는대로 결과를 내보낼 뿐이다. 컴퓨터는 현재의 자료에 대한 분석법이 옳은지 그른지 판단하지 못한다. 패키지 프로그램을 사용하려는 이는 프로그램을 작성하기 전에 자신이 선택한 분석법이 과연 적절하였는지 한번 더 숙고해야 한다. 패키지 프로그램에서 백여 항목의 통계량을 출력하였더라도 쓸모있는 것은 단 한두개에 불과한 경우도 많이 있다. 요컨대 컴퓨터가 대행한 계산 결과를 해석하는 것은 분석자의 소임임을 유념하라. 모든 분석결과는 평범한 어휘로 통계에 문외한들에게도 쉽게 설명할 수 있어야 한다.

※ 생각해보기

① 기상예보에서 '내일 서울에 비올 확률이 50%'라고 하였다. 얼마나 가치있는 진술일까?

② 동전 던지기를 100번 하였더니 앞면이 90번, 뒷면이 10번 나왔다. 다시 한번 던지기 전에 세 사람이 각각 다음과 같은 주장을 하였다. 누구의 주장이 합리적일까?

a) 앞뒷면이 각각 50번 정도씩 나와야하는데 그 동안 뒷면이 너무 안 나왔다. 따라서 나는 뒷면에 1억을 걸겠다.

b) 그간 앞면이 지나치게 많이 나온 것으로 보아 이 동전은 분명히 앞면이 나올 확률이 매우 큰 동전일 것이다. 따라서 나는 앞면에 1억을 걸겠다.

c) 동전은 과거를 기억하지 못한다. 사람이 기억할 뿐이지. 따라서 101번째의 결과는 앞이나 뒤가 나올 확률이 반반일텐데… 나는 도박에 참여하지 않겠다.

※ SAS 시스템 소개

오늘날 통계학에서 흔히 쓰이는 자료분석법은 헤아릴 수 없을만큼 많다. 가장 초보적인 자료분석법의 예를 들자면, 수집된 자료를 기초로 히스토그램 따위의 도표 그리기, 산술평균, 중앙값, 표준편차들과 같은 간단한 기술통계량의 계산 등을 꼽을 수 있고, 선형모형을 가정한 응용통계기법 중 많이 쓰이는 것으로는 회귀분석, 분산분석 등을 들 수 있다. 상황에 따라 적용할 수 있는 자료분석기법이 많음에 버금가게, 통계조사 혹은 통계실험으로부터 얻어지는 자료의 구조 또한 천태만상이다. 따라서, 각 경우마다 개별적인 컴퓨터 프로그램을 작성하여 분석을 해야한다면 지극히 불편할 것이다. 이러한 단점을 극복하기 위하여 등장한

개념이 패키지 프로그램(package program), 또는 소프트웨어 패키지(software package)로, 패키지 프로그램은 많은 프로그램이 동일한 구조 하에 통합된 소프트웨어이다.

일반적으로 어떠한 패키지 프로그램이든 한가지 공통된 작업 형태는, 그 자체의 고유한 입력 방법으로 주어진 자료를 읽어들이고 사용자가 선택한 단위 프로그램으로 자료 처리를 한 후, 처리 결과를 그 나름의 표준화된 방식대로 출력하는 것이다. 그리고, 패키지 프로그램을 구성하는 하나의 모듈(module)에서 만들어진 데이타는 모든 모듈에서 공유할 수 있도록 설계된다.

특히, 통계자료처리와 분석이 주목적인 패키지 프로그램을 통계 패키지 프로그램이라 하며, 통계 패키지 프로그램 중 현재 가장 널리 사용되고 있는 것이 SAS이다. SAS 이외에도 S-PLUS, SYSTAT, NCSS, STATISTICA, BMDP, MINITAB, SPSS같은 통계 패키지 프로그램이 수없이 많으나, 지원되는 통계 분석법의 종류라든가 인기도 등에서 SAS는 타 프로그램의 추종을 불허하고 있다.

SAS는 범용 통계 패키지 프로그램이다. 이 말은 SAS 하나로 거의 모든 통계 분석을 할 수 있음을 뜻한다. 이와 대치되는 개념이 전문 통계 패키지 프로그램으로, 이런 소프트웨어에서는 특정 분야의 자료 처리만 전문적으로 수행할 수 있다. 예를 들어, SCA와 같은 패키지 프로그램은 시계열분석만을 전문적으로 수행하며, ECHIP은 주로 실험계획에 관련된 소프트웨어이다. 전문 통계 패키지 프로그램들은 그 분야에 대한 자료처리와 분석에는 탁월하지만, 똑같은 데이타로 색다른 분석을 시도하지 못하는 단점이 있다. 반면에, 하나의 시스템으로 모든 자료 처리를 일관성있게 수행할 수 있고, 종류가 다른 통계 분석 작업을 같은 환경에서 처리할 수 있는 것이 범용 소프트웨어의 장점이다.

SAS는 Statistical Analysis System 또는 Strategic Applications Software를 의미한다. 흔히 '쌔쓰', 또는 영어 자모 그대로 '에쓰에이에쓰'라 읽는다. SAS는 SAS 연구소(SAS Institute)에서 개발한 **통계 패키지 프로그램**(statistical package program)이다. 가장 최신판 PC/SAS는 6.11 판으로 Windows 3.1 또는 Windows 95 하에서 구동된다. DOS용 최신판은 6.04 판이다. 6.11 판의 총 크기는 약 450 MB로 지나치게 방대한 느낌이다. 이중 순수하게 통계 분석용 소프트웨어는 약 250 MB 정도를 차지한다. DOS용 판은 약 15~40 MB의 하드 디스크 공간이 요구된다.

2. 기술통계분석

SAS에서 기술통계 분석은 UNIVARIATE 절차와 MEANS 절차에서 담당한다.

2.1 평균, 분산, 자유도

수집된 n 개의 관측(observation)을 $x_1,\cdots,x_n$으로 표기하자. 이 n 개의 관측 모두를 통틀어 자료(data)라 한다. 그리고, 각 관측에 대하여 실제로 얻은 값을 관측값(observed value)이라 한다.

(1) 평균

평균(mean, 또는, average)은 관측값들의 합을 관측들의 총수로 나눈 값이다. 수식으로 표현하면 다음과 같다. 평균은 자료의 균형점, 즉 중심(center)을 하나의 숫자로 요약하는 통계량이다.

$$\text{평균} \quad \bar{x} = \sum_{i=1}^{n} x_i \,/ n$$

(2) 분산과 표준편차

분산(variance)과 표준편차(standard deviation)는 평균을 기준점으로 자료들이 얼마나 좌우로 퍼져있는지를 재는 측도이다. 이런 이유로 평균을 중심 경향(central tendency)의 측도라 하고, 분산과 표준편차는 퍼짐(dispersion)의 측도라 한다. 평균을 기준으로 자료에 포함된 관측값들의 퍼짐성을 재려면, 각 관측값들이 평균으로부터 떨어진 거리, 즉, 편차(deviation)를 구해야 한다. 따라서, 편차는 자연스럽게 「관측값-평균」으로 정의된다.

$$\text{분산} \quad s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / (n-1)$$

따라서, 분산은 편차 제곱합들의 평균이라 할 수 있다. 여기서 분모의 $(n-1)$을 자유도(df: degree of freedom)라 한다. 표준편차는 분산의 제곱근이다. 즉, 표준편차 $s = \sqrt{s^2}$. 대부분의 경우 분산보다 표준편차를 더 많이 쓰는데 그 이유는 평균의 단위와 표준편차의 단위가 같기 때문이다.

자유도란 무엇인가? 자유도는 데이타에서 상호 의존하지 않는 관측의 갯수이다. 예를 들어 설명한다. 다음 다섯개 값을 관측하였다.

5 2 4 1 3

이 자료에 대한 자유도는 5이다. 이 자료에서 평균을 계산하였더니 그 값이 3이었다. 그 후에 실수로 원래 관측 하나를 잃어버렸다고 가정하자. 예를 들어 4를 잃었다고 하자. 아래에서 빈칸 □는 그 관측이 무슨 값인지 모름을 뜻한다.

5 2 □ 1 3

이 지워진 값이 얼마인지 알아낼 수 있을까? 이것은 아주 간단하다. 평균이 3이었고 총 관측수는 5라는 정보에서 자료에 포함된 점수들의 합은 15가 됨은 쉽사리 알 수 있다. 그리고, 이 15에서 잃어버리지 않은 나머지 네 개의 관측값을 빼면 곧바로 4가 나온다. 두 개의 점수를 잃어버려도 복원이 가능할까? 이는 절대 불가능하다.

여기서 알 수 있는 사실은 「일단 평균을 계산하면, 원래 자료 중의 한 값을 잃어도 전혀 정보의 손실이 없다」는 평범한 진리이다. 즉, 평균값을 알고있을 때 자료 중의 어느 한 값은 언제나 잉여 정보(redundant information)가 되며, 따라서 어느 하나의 관측값을 모르더라도 전체 자료가 함축하고 있는 정보에 손실이 가는 것은 아니다. 그리고, 분산은 먼저 평균을 구한 연후에 계산이 가능하므로, 분산의 자유도는 언제나 $(n-1)$, 즉, 「관측수-1」이 된다.

참고로, 평균을 계산할 때는 관측들의 합을 자유도 n으로 나누고, 분산을 계산할 때는 편차들의 제곱합을 자유도 $(n-1)$로 나눈다고 생각해도 된다.

실험을 한 뒤 얻은 데이타에 대한 보고서나 연구 논문을 작성할 때 데이타의 변동(variability)에 관련된 정보를 제공하는 것은 상식적인 절차이다. 수집된 데이타의 중심은 흔히 평균(average, mean)으로 나타낸다. 그리고, 모든 데이타 값들이 데이타의 중심으로 선택된 평균으로부터 얼마만큼 퍼져있는지는 주로 표준편차(standard deviation), 또는 표준오차(standard error)로 정량화한다.

표준오차 $se = s / \sqrt{n}$

수집된 데이타에서 평균 중심의 변동을 보고하는 형식은 천차만별이지만 일반적으로 추천할 만한 형태는 다음 세 가지이다.

① $\overline{X} \pm s$

② $\overline{X} \pm se$

③ $\overline{X} \pm 2 \times se$

이때 표본 크기와 측정 단위를 항상 명시해야한다. ①과 같이 표준편차를 보고하는 것보다

는 ②, ③과 같이 표준오차를 보고하는 편이 더 낫다. 표준오차는 평균값의 정밀도를 대표하는 값이기 때문이다.,③에서는 표준오차의 2 배를 보고하는데, 이 형식은 통계 추론에서 언급되는 95% 신뢰구간의 계산과 동일하다. 어느 경우든 표준편차나 표준오차의 값이 작으면 데이타가 평균 근처에 많이 밀집되어 있고, 따라서 측정이 정밀하였다고 볼 수 있다. 그러나, 이같은 진술은 상대적이기 때문에 비슷한 실험 간의 정밀도를 비교할 때나 사용할 수 있다.

오렌지 쥬스	신맛: $\overline{X} \pm se$ (n)
A	7.583 ± 0.288 (12)
B	6.667 ± 0.256 (12)
C	6.917 ± 0.288 (12)
D	5.583 ± 0.484 (12)

이와 같이 보고된 통계량들을 보면, 오렌지 쥬스 A가 신맛이 가장 강하고 오렌지 쥬스 D의 평가 점수가 다른 오렌지 쥬스들에 비하여 변동이 심함을 쉽게 알 수 있다.

(3) 줄기와 잎 그림, 그리고 상자그림

줄기와 잎 그림(stem-and-leaf plot)은 최근 급속히 발전되고 있는 탐색자료분석(exploratory data analysis) 분야에서 개발된 자료기술기법이다. 줄기와 잎 그림은 상당히 잘 고안된 그래프 기법으로, 새로운 유형의 도수분포도로 간주할 수 있다. 줄기와 잎 그림은 종래의 도수분포도에 비하여, 도수, 데이타 값들의 집중도, 분포의 모양, 그리고 실제 데이타 값 등 전반적인 자료의 변화 추이를 쉽게 알 수 있게 한다.

다음 90 개의 데이타 값들을 이용하여 줄기와 잎 그림을 어떻게 그리는지 살펴보자.

843	560	792	668	637	707	522	953	592	752
621	991	503	719	495	404	370	184	470	832
918	550	437	499	272	656	802	1068	830	448
270	171	858	336	585	521	794	860	444	411
201	453	250	112	322	346	212	758	306	1000
737	638	305	527	592	436	607	265	482	409
818	441	451	640	351	925	322	869	1148	572
570	618	480	792	709	766	574	390	629	1005
514	559	289	439	739	578	494	616	487	548

이 90 개의 데이타 값들은 전부 세 자리 아니면 네 자리의 숫자들이다. 각 데이타 값들에서 백단위 이상의 숫자들(세 자리 데이타 값에서는 첫번째 숫자)을 줄기로 생각하고, 나머지 십자리 이하의 두 숫자를 잎으로 생각하자.

위의 자료에서 최소값은 112, 최대값은 1148이므로, 줄기에 해당하는 숫자는 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11의 11 개이다. 도수표와 마찬가지로, 줄기에 해당하는 값이 한 행이 되며, 각 데이타 값에 대응되는 줄기를 찾아 잎에 해당하는 숫자를 옆에 쓴다. 줄기와 잎을 구분하기 위하여 보편적으로 수직선을 긋는다.

다음 그림이 위 데이타 값들에 대한 줄기와 잎 그림이다. 여기서, 각 줄기 내의 잎들은 크기에 따라 순서화한 것이다.

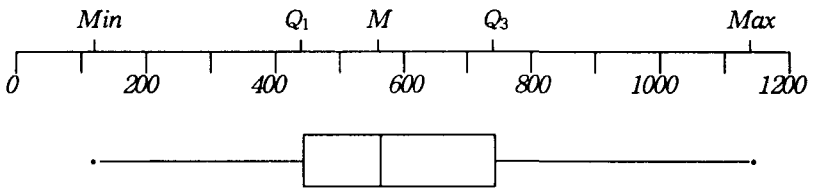
1	12 71 84
2	01 12 50 65 70 72 89
3	05 06 22 22 36 46 51 70 90
4	04 09 11 36 37 39 41 44 48 51 53 70 80 82 87 94 95 99
5	03 14 21 22 27 48 50 59 60 70 72 74 78 85 92 92
6	07 16 18 21 29 37 38 40 56 68
7	07 09 19 37 39 52 58 66 92 92 94
8	02 18 30 32 43 58 60 69
9	18 25 53 91
10	00 05 68
11	48

줄기와 잎을 결정하는 가장 좋은 방법은 없다. 줄기를 달리해서 몇 가지 그려보고 어느 것이 가장 그럴듯해 보이는지 직관적으로 결정해야한다. 예를 들어, 점수들이 123456과 같은 여섯 자리의 숫자이면, 처음 두 숫자를 줄기로, 가운데 둘은 잎으로, 나머지 끝의 두 숫자는 무시한다. SAS에서는 출력 공간을 감안하여 나름대로 최적이라 판단되는 줄기와 잎 그림을 출력한다.

상자 그림(box plot)은 줄기와 잎 그림에 함축된 정보를 토대로 데이타의 중심과 퍼짐에 대한 몇 가지 측도를 병합하여 만들며, 주로 분포의 대칭성, 분포의 양 꼬리 부분의 집중도 등을 탐색할 때 많이 사용된다. 상자 그림은 자료의 최소값, 아래 사분위값, 중앙값, 위 사분위값, 최대값 등 다섯 개의 분위수들을 기초로 그린다.

위의 자료에서, 최소값은 112, 아래 사분위값은 436, 중앙값은 559.5, 위 사분위값은 739, 최대값은 1148이며, 이 다섯 가지의 값이 상자 그림을 만드는 기본적인 측도들이다. 상자 그림은 아래 사분위값과 위 사분위값을 네모난 상자로 연결한 다음 중앙값에 선을 긋고, 그 다

음 최대값과 최소값을 점으로 표시한 다음 점과 상자 사이를선으로 연결하여 완성한다.
위 자료에 대한 상자 그림은 다음과 같이 그린다.



최대값과 최소값을 상자와 연결하는 선을 구레나룻(whisker)이라 부른다. 지금 소개한 상자 그림은 간단한 형태의 상자 그림이지만, 위와 같은 상자 그림에서 다음과 같은 정보를 쉽게 읽을 수 있다.

① 아래 사분위값과 위 사분위값
② 사분위 범위
③ 가장 큰 값과 가장 작은 값
④ 자료 분포의 대칭성, 또는 비대칭성

그리고 중앙값이 아래 사분위값에 치우쳐있으므로 중앙값 아래쪽으로 자료들의 집중도가 크다고 판단할 수 있으며, 왼쪽의 구레나룻이 오른쪽보다 짧으므로 분포의 형태가 왼쪽은 급경사, 오른쪽은 완만한 경사를 이루는 비대칭 분포임을 알 수 있다.

그리고, 이 상자 그림을 이용하여 나머지 자료와 동떨어진 값을 갖는 관측인 이상점(outlier)들을 찾기도 하는데, 상자의 양끝에서 각각 1.5 배의 사분위범위(IQR=Q3−Q1) 밖에 위치한 관측을 이상점으로 판정한다. 특히 3 배의 사분위범위 밖에 위치한 관측들은 극단적인 이상점들(extreme outliers)이다.

Schematic plot은 비슷한 속성을 갖는 여러 자료를 동시에 비교하고자 할 때 사용되는 기법으로, 자료마다 상자 그림을 동일 축척하에 나란히 그려 분포 형태 등을 한 눈에 볼 수 있도록 하고 있다.

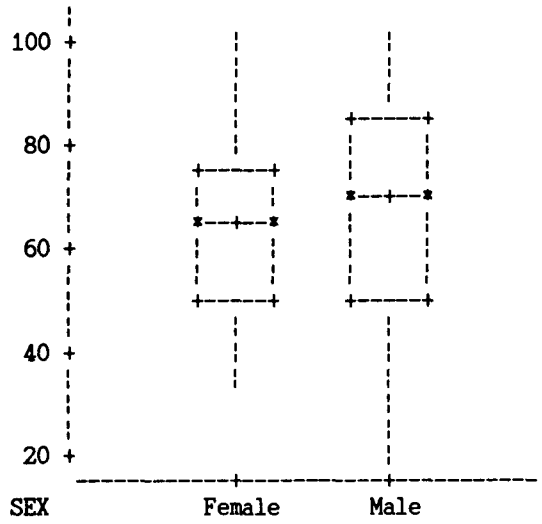
```
                          The SAS System

                        Univariate Procedure

Variable=SCORE

                              Moments

          ==>  N                 89  Sum Wgts         89
          ==>  Mean        55.75281  Sum            4962
           =>  Std Dev     24.08804  Variance   580.2337  <=
               Skewness    -0.52298  Kurtosis   -0.41141
               USS           327706  CSS        51060.56
               CV           43.20507  Std Mean   2.553327  <==
               T:Mean=0     21.83536  Pr>|T|       0.0001


                         Quantiles(Def=5)

                 100% Max        93       99%        93
          ==>     75% Q3         75       95%        89
          ==>     50% Med        55       90%        87
          ==>     25% Q1         40       10%        20
                   0% Min         0        5%         6
                                            1%         0
                 Range           93
          ==>    Q3-Q1           35
                 Mode            38
```

```
          Stem Leaf                     #             Boxplot
             9 13                       2                |
             8 5667778999              10                |
             8 000134                   6                |
             7 556778                   6             +-----+
             7 0001224                  7             |     |
             6 5588                     4             |     |
             6 02333344                 8             |     |
             5 558                      3             *--+--*
             5 00111222344             11             |     |
             4 577889                   6             |     |
             4 0044                     4             +-----+
             3 5688889                  7                |
             3 133                      3                |
             2 6                        1                |
             2 011                      3                |
             1 9                        1                |
             1 3                        1                |
             0 667                      3                |
             0 004                      3                |
               ----+----+----+----+
          Multiply Stem.Leaf by 10**+1
```

The SAS System

Univariate Procedure
Schematic Plots

Variable=SCORE1



3. 설문지 분석

설문지(questionnaire, enqete)를 이용한 통계조사 연구는 신문지상에서 흔히 접할 수 있다. 이 장에서는 이런 분석에 필수불가결인 카이제곱(χ^2) 검증을 공부한다. 이와같은 유형의 자료는 **교차표**(crosstable) 형식으로 정리되며, 교차표로 정리된 자료의 분석은 더 일반적으로 **범주형 자료분석**(categorical data analysis)이라 부른다. SAS에서는 FREQ 절차에서 담당한다.

3.1 독립성 검증에 대한 기초 개념

범주형 자료분석은 **분할표**(contigency table), 곧, 이원 교차표(two-way crosstable) 형식으로 정리된 자료를 분석하는 응용 통계학의 한 기법이다. FREQ 절차에서는 각 분할표마다 독립적으로 검증통계량과 연관측도들을 출력하며, 이러한 통계량들은 대부분 열변수와 행변수 간의 독립성을 검증할 때 쓰이는 χ^2 통계량의 개념에 기초를 둔다.

한 가지 예를 들어 보자. 다음의 2×2 표는 남녀별로 사탕을 자주 먹는지 조사한 결과를 요약한 것이다.

관측도수	자주 먹는다	자주 먹지않는다	합계
남자	10	50	60
여자	70	20	90
합계	80	70	150

이 데이타에 대하여 독립성 검증을 한다함은 남자냐 여자냐에 따라 사탕에 대한 선호도에 차이가 있는지를 검증하는 것을 뜻한다. 위 데이타를 보면, 남자에 대한 행의 데이타 값은 10에서 50으로 증가하고 있는 반면, 여자에 대한 행의 데이타 값은 70에서 20으로 감소하고 있다. 즉, 직관적으로 성별에 따라 사탕에 대한 선호도가 다름을 알 수 있다. 위 데이타는, 아직 통계적 검증을 하지는 않았지만, 성별과 사탕의 선호도 사이에는 의존성이 있음을 보여주고 있다. 즉, **성별과 사탕은 서로 독립이 아님**을 시사한다.

이러한 독립성 검증의 통계적 가설은 다음과 같이 세울 수 있다.

H_0: 두 변수는 서로 독립이다.
H_1: 두 변수는 서로 독립이 아니다.

그렇다면, 이러한 가설을 어떻게 검증할 수 있을까? 만일 좀전의 데이타가 두 변수 간

의 완벽한 독립성, 즉, 사탕을 좋아하는데는 남녀 구분이 없다는 사실을 뒷받침한다면 각 셀(cell)의 도수가 어떻게 분포되어야 하는지 생각하여 보자. 그러기 위하여, 개별적인 셀 도수가 없는 다음과 같은 표를 고려한다.

관측도수	자주 먹는다	자주 먹지않는다	합계
남자 여자			60 90
합계	80	70	150

독립성 검증의 아이디어는 간단하다. 즉, 두 변수가 서로 완전히 독립이라는 가정 하에 이론적인 기대도수(expected frequency)를 산출한 뒤, 실제 관측도수(observed frequency)와 기대도수 간에 차이가 많이 나면 독립이 아니라고 결론짓고, 차이가 크지 않으면 서로 독립이라 한다. 이때, 차이가 많이 났느냐 아니냐는 카이제곱 분포로 판정한다.

성별과 사탕이 서로 독립이라고 할 때 각 셀의 기대도수를 어떻게 결정하는지 살펴보자. 우선 사탕을 자주 먹는 남자들이 몇 명일지 생각해보자. 전체 조사 인원이 150 명이고, 남녀를 불문하고 사탕을 자주 먹는 사람이 80 명이다. 이때, 이 조사에서 남자는 전부 60 명이므로, 남자 중에 사탕을 자주 먹는 사람은 다음 비례 관계로 결정된다.

$$150 : 80 = 60 : ?$$

따라서, 이 셀의 기대도수는 상호독립이란 가정 하에 다음과 같이 계산된다.

$$(80 \times 60) / 150 = 32$$

나머지 셀들의 기대도수도 똑같은 방식으로 결정할 수 있으나, 현재 다루는 교차표가 2×2 표임을 감안하면 더 쉽게 구하는 방법이 있다. 즉, 남자로서 사탕을 자주 먹지않는 사람들에 대한 기대도수는 남자의 총수가 60이므로, 60−32=28이 되며, 여자에 대한 셀들의 기대도수도 비슷하게 구하면 된다. 다음은 이렇게 구한 기대도수표이다.

기대도수	자주 먹는다	자주 먹지않는다	합계
남자 여자	32 48	28 42	60 90
합계	80	70	150

독립성 검증에서는 관측도수와 기대도수를 비교하여, 도수들 간에 많은 차이를 보이면 독립이 아닌거고, 도수들 간에 그다지 큰 차이가 없을 때는 서로 독립이라고 판단한다. 따라서, 검증 통계량이 무엇이든, 그 공식은 「관측도수−기대도수」의 함수로 주어지게

된다. 독립성 검증시 한 셀에 대한 편차만을 고려하는 검증하는 것은 무의미하므로 모든 셀에 대하여 이러한 편차들의 합을 계산하여 독립성 유무를 검증하게 되는 것은 상식적인 절차이다.

만일 위 사탕 자료에 대하여 독립성 검증을 하면 대응되는 유의확률이 거의 0으로 출력된다. 따라서, 성과 사탕 사이에는 의존성이 있다고 말할 수 있다. 데이타를 좀더 잘 살펴보면, 남자에 비하여 여자가 사탕을 더 자주 먹는다고 할 수 있겠다.

3.2 유의확률(significance probability)

통계적 가설검증(statistical hypotheses testing)은 데이타를 기초로 실험자가 주장하고 싶은 가설의 옳고 그름을 판정하는 통계 절차이다. 통계적 가설검증에서는 언제나 귀무가설(null hypothesis)과 대립가설(alternative hypothesis)의 두 가설을 설정하여 데이타에 담긴 정보를 토대로 두 가설 중 어느 하나로 결론을 내리는 형식을 취한다. 앞으로는 귀무가설을 H_0, 대립가설을 H_1으로 표기한다. 이때 실험자가 주장하려는 사실은 대부분의 경우 대립가설로 설정한다.

예를 들어, 오렌지 쥬스들 간에 신맛의 차이가 있다는 사실을 보이려면 대응되는 귀무가설과 대립가설은 다음과 같이 써야 한다.

H_0 : 오렌지 쥬스들 간에 신맛의 차이는 없다.
H_1 : 오렌지 쥬스들 간에 신맛의 차이가 있다.

일반적으로 실험자가 데이타를 통해 새로 보이고 싶은 주장은 언제나 대립가설에 설정한다. 따라서, 실험 결과를 통계분석할 때 실험자는 귀무가설을 기각(reject)하고 대립가설을 채택(accept)하기를 바라게 된다. 통계적 가설검증은 보수적인 관점에서 개발되었기 때문에 데이타를 통한 확실한 증거가 없이는 함부로 처리들 간에 차이가 있다는 사실을 받아들이지 않는다. 따라서, 귀무가설을 쓸 때는 언제나 '처리들 간에 차이가 없다', '처리들 간에는 아무 관계도 없다'는 식의 진술을 하게 된다. 수리적으로 표현하면 귀무가설에는 반드시 = 부호가 있게됨을 명심하라.

통계적 결정은 데이타에 담긴 정보를 기초로 한다. 따라서 실험자나 연구자가 바라지는 않지만 언제나 판단 착오를 할 위험이 있다. 즉, 통계적 결정에는 귀무가설이 옳은데도 불구하고 귀무가설을 기각할 오류와 귀무가설이 거짓임에도 불구하고 귀무가설을 채택할 오류의 두 가지 오류가 수반된다. 전자를 제1종 오류(type I error), 후자를 제2종 오류(type II error)라 부른다. 또한, 제1종 오류를 범할 확률을 α, 제2종 오류를 범할 확률을 β라 표기한다. 다음 표에 α와 β 간의 관계를 요약하였다.

통계적 결정	귀무가설 H_0	
	참	거짓
H_0를 기각함	제1종 오류 α	옳은 결정 $1-\beta$
H_0를 채택함	옳은 결정 $1-\alpha$	제2종 오류 β

α와 β의 두 확률은 통계적 결정을 잘못내릴 확률이므로 가능한한 작을수록 좋다. 그러나, 이론적으로 두 오류의 확률을 동시에 작게 만들수는 없으므로 일단 α의 값을 0.01, 0.05 등과 같은 작은 값으로 고정시킨 다음 β를 최소화하는 방식을 취하여 통계적 가설검증법을 만든다. 이렇게 선택된 α의 값을 유의수준(significance level)이라 한다. 일반적으로 유의수준은 0.05로 많이 잡는다. 이 말은 귀무가설이 옳은데도 불구하고 데이타가 잘못 나와서 대립가설을 채택할 오류의 확률이 5/100라는 의미이다.

반복하지만, 유의수준이 0.05라는 말은, 표본을 100번 독립적으로 얻어 똑같은 검증을 할 때 다섯번 정도는 귀무가설이 맞는데도 귀무가설을 기각할 오류를 범할 수 있다는 의미이다. 좀더 안전한 결론을 내리고 싶으면 유의수준을 0.01로 설정할 수 있다. 물론 유의수준이 0.01이라는 말은 100번에 한 번 정도 귀무가설이 옳은데도 불구하고 귀무가설을 기각하는 오류를 범할 수 있다는 뜻이다.

그러나 유의수준 α의 값을 아주 작게 설정한다고 반드시 가설검증이 정확해지는 것도 아니며, 그렇다고 아주 크게 한다고해서 정확도가 떨어지는 것도 아니다. 예를 들어, α를 0.5로 놓으면 0.05로 놓을 때에 비하여 귀무가설이 옳음에도 불구하고 귀무가설을 기각할 확률이 매우 커진다. 반대로 α를 0.0001로 놓으면 0.05로 놓을 때에 비하여 귀무가설이 옳지 않음에도 불구하고 대립가설을 채택할 확률이 아주 작게된다.

일반적으로 물리 및 화학 실험 자료와 같이 정교한 데이타일 경우에는 유의수준을 0.05, 0.01로 놓게 되지만, 사회 및 경제 분야의 자료와 같이 거친 데이타를 분석할 때는 유의수준을 0.1, 0.2로 놓게 되기도 한다.

이와 같은 유의수준 선택에 대한 다양성 때문에 특히 통계 소프트웨어들에서는 언제나 유의확률(significance probability)을 출력하여 실험자에게 가설 검증의 결과 판정을 맡기고 있다. 유의확률을 p 값(p-value)이라 부르기도 한다. 유의확률은 데이타를 기초로 귀무가설이 기각될 유의수준을 역산한 확률값이다. 즉, **현재 갖고있는 데이타에 대해서 가설 검증을 할 때 애초에 유의수준을 얼마로 잡았어야 귀무가설을 기각할 수 있는지를 계산한 값이 유의확률이다.**

따라서, 유의수준을 0.05로 정했을 때 계산된 유의확률이 0.05보다 작게 나와야 귀무가설을 기각하고 대립가설을 채택할 수 있으며, 반대로 유의확률이 0.05보다 크게 나오면 대립가설

을 채택할 수 없다. 따라서, 일반적으로 대립가설을 채택하는 것이 목표인 실험자들은 유의확률값이 0.05보다 작게 나오기를 바라게 된다. 특히 유의확률값이 0.01보다도 작게 나오면 아주 확실하게 귀무가설을 기각하고 대립가설을 결정지을 수 있다. 가설검증 절차를 포함하는 많은 SAS 절차에서는 **유의확률**(significance probability)을 출력하여 사용자로 하여금 가설검증의 결과 해석에 도움을 주고 있다.

3.3 가설검증과 신뢰구간

어떤 전구 회사에서 제조하는 100 와트 전구들의 평균 수명이 지금까지는 1000 시간이었다고 가정하자. 새로운 공법을 도입하여 전구를 생산하였다. 새로운 전구들의 평균 수명 μ가 종래 제품의 평균 수명보다 더 길 것으로 기대되는데, 정말 그러한지 소량의 표본을 추출하여 검증하고 싶다. 이런 경우, 귀무가설과 대립가설은 다음과 같이 설정할 수 있다.

$$H_0: \mu = 1000$$
$$H_1: \mu > 1000$$

새로운 전구를 10 개 선택하여 수명을 재었더니 평균 수명이 2000 시간이 나왔다고 하면, 우리는 쉽사리 귀무가설을 **기각**(reject)하고, 대립가설을 **채택**(accept)할 수 있다. 즉, 새로운 제품이 종래의 제품보다 수명이 더 길다고 할 수 있을 것이다. 그러나, 표본으로 추출된 새로운 전구들의 평균 수명이 1040 시간으로 나왔다면, 이때도 귀무가설을 기각할 수 있을까?

평균 수명 1040 시간도 사실은 신제품이나 구제품이나 수명에 별 차이가 없음에도 우연히 다소 수명이 긴 전구들이 추출되어 나온 결과일 수도 있다. 그렇다면, 도대체 새로운 전구의 평균 수명이 얼마로 나와야 귀무가설을 기각할 수 있을까? 이렇게, 귀무가설의 선택과 대립가설의 선택의 갈림점을 **임계점**(critical point)라 하며, 임계점의 결정은 표본 통계량의 확률분포로부터 결정된다.

통계학에서 표본조사를 하는 이유는 표본이 추출된 모집단을 규정하는 모수를 추정하기 위함이다. 예를 들어, 모집단의 평균 μ를 표본평균으로 추정한다면, 이런 추정 방법을 **점추정**(point estimation)이라 한다. 이 이유는 미지의 모수 μ를 표본평균이라는 하나의 숫자로, 즉, 실수선 상의 한 점으로 추정하기 때문이다.

그러나 자료에서 계산된 단 한 점으로 미지의 모수를 추정하는데는 상당한 무리가 따르기 때문에, 모수의 추정값로서 단일 숫자를 사용하기보다는 모수가 포함될만한 숫자들의 범위를 명시하는 추정 방법이 더 보편적이다. 이와같은 추정 방법을 **구간추정**(interval estimation)이라 한다.

통계학에서 미지의 모수 θ에 대한 구간 추정은 일반적으로 다음과 같은 공식으로 주어진다.

δ ± (δ의 표준오차) × c

여기서, δ는 표본에서 산출된 θ의 추정값이다. 그리고, c는 δ의 통계 분포로 결정되는 상수이다.

다시 말해서, 어떤 주어진 모수에 대한 구간추정이란, 표본에서 산출된 그 모수에 대한 점 추정값을 중심으로, 점 추정량에 대한 표준 오차의 상수배만큼 좌우로 벌어진 구간을 구하는 추정 방법이다.

또, 이런 식으로 결정된 구간을 **신뢰구간**(confidence interval)이라 부른다. 그리고, 신뢰구간의 왼쪽 끝을 **신뢰하한**(lower confidence limit), 오른쪽 끝은 **신뢰상한**(upper confidence limit)이라 한다. 많은 SAS 절차에서 구간추정에 관한 통계량을 출력하며, 직접 신뢰구간을 출력하지 않는 경우에는 추정량의 표준오차나 근사표준오차를 출력하여 신뢰구간을 쉽게 구할 수 있도록 하고 있다.

한국인의 평균키를 추정하고자 할 때 신뢰구간으로 (1 m, 2 m)를 제시한다면, 분명히 이 구간 내에 평균키에 해당하는 점이 포함될 것은 100% 확실한 일이다. 이런 구간을 100% 신뢰구간이라 부를 수 있겠다. 그러나, 이런 신뢰 구간은 아무짝에도 쓸모가 없다. 다소 틀릴 확률이 있더라도 신뢰구간의 범위를 짧게 지정해야 좋음은 명약관화하다. 통계학에서는 주로 95% 신뢰구간, 99% 신뢰구간 등을 사용한다. 95% 신뢰구간이란, 동일한 모집단에서 표본을 추출하여 신뢰구간을 산출하는 작업을 100 회 반복한다고 가정할 때, 100 개의 신뢰구간들 중에서 적어도 95 개의 신뢰구간들은 미지의 모수를 포함한다는 것을 뜻한다. 물론 기껏해야 다섯 개 정도의 신뢰구간에는 미지의 모수가 포함되지 않을 위험이 있다.

여기서 95%나 99%와 같은 숫자를 **신뢰수준**(confidence level), 또는 **신뢰계수**(confidence coefficient)라 하며, 신뢰수준의 퍼센티지를 100%에서 뺀 값, 즉, 5%나 1% 따위가 **유의수준**(significance level)이다.

3.4 분할표에 대한 χ^2 검증

χ^2 검증법은 범주형 자료(categorical data)에 대한 통계적 검증 방법으로 다음과 같은 두 가지 유형이 있다.

독립성(independence) 검증
동질성(homogeneity) 검증

이들은 이차원 이상의 도수표, 즉 분할표(contingency table) 자료에 대한 것이다. 독립성 검증과 동질성 검증 간에 개념 차이는 있으나 동일한 분석법을 공유한다.

분할표는 두 요인에 대한 범주화 자료를 요약한 이차원 도수표이다. 이때 두 요인을 각각

행요인(row factor)과 열요인(column factor)이라 부른다. 행요인에는 범주가 r 개, 열요인에는 범주가 c 개 있다고 하자. 분할표 자료의 구조는 다음과 같이 정리할 수 있다.

행 \ 열	1	2	…	c	합계
1	f_{11}	f_{12}	…	f_{1c}	R_1
2	f_{21}	f_{22}	…	f_{2c}	R_2
⋮	⋮	⋮	⋱	⋮	⋮
r	f_{r1}	f_{r2}	…	f_{rc}	R_r
합계	C_1	C_2	…	C_c	n

이러한 $r \times c$ 분할표에 대한 검증에는 독립성 검증과 동질성 검증이 있다 하였다. 독립성 검증은 두 요인 간에 상호 연관성이 있어 서로 영향을 끼치는지를 조사할 때 사용한다. 이때의 귀무가설은 '두 요인은 서로 독립이다', 또는 '두 요인 간에 아무 관련이 없다'이다. 동질성 검증은 하나의 요인의 각 범주마다 도수의 분포가 동일한지 여부를 판정할 때 사용한다. 즉, 행요인의 첫번째 범주에서 열요인의 범주에 따른 도수 분포의 변화 패턴이 행요인의 두번째, 세번째, …, r번째 범주들에서의 변화 패턴들과 같다고 할 수 있는지를 검증하는 것이다. 동질성 검증에서는 어느 하나의 요인을 고정시키는 점이 독립성 검증과 다르다. 그러나, 개념 차이는 있지만 분석법은 같다.

분할표에서 R_i들은 각 행의 합계, C_j는 각 열의 합계, n은 총합계이다. 이때 기대도수를 구하는 일반 공식은 다음과 같다.

$$e_{ij} = \frac{R_i \times C_j}{n}$$

분할표에 대한 검증 통계량은 다음과 같이 주어진다.

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

이 통계량은 자유도 $(r-1)(c-1)$의 χ^2 분포를 따르며, 이 χ^2 통계량이 범주형 자료분석의 근간이 된다. 만일 두 분 변수가 서로 독립이 아니라면, 각 셀의 편차가 상당히 크게 나타나게 되어 χ^2 통계량의 값도 커지며 출력되는 유의확률의 값이 0에 가까와진다. 보통 이 유의확률이 값이 0.05 이하일 때, 우리는 두 분류 변수가 독립이 아니라고 결론짓는다.

만일 $r=c=2$라면 다음과 같이 연속성 수정된 검증 통계량을 사용한다.

$$\chi_c^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(|f_{ij} - e_{ij}| - 0.5)^2}{e_{ij}}$$

[예] 한 식품 회사에서 신제품을 시판하기 전에 소비자의 반응을 보기 위하여 선택된 사람들에게 시식을 하게 한 다음 설문 조사를 하였다. 다음은 그 가운데 일부의 데이타로서, 직업 수준에 따라 구매 의사를 확인한 분할표 자료이다.

구 매 의 사	직업			합계
	사무직	육체 노동자	기타	
반드시 사겠다	87	54	22	163
아마 살 것이다	57	22	15	94
살 수도 안 살 수도 있다	19	8	4	31
아마 사지 않을 것이다	2	2	1	5
절대로 사지 않겠다	1	0	0	1
모르겠다	3	0	0	3
합계	169	86	42	297

이 자료에 대해서 우리가 알고 싶은 것은 직업 수준과 구매 의사 간에 어떤 연관성이 있는지 여부이다. 이와 같은 분할표 자료에 대한 검증을 할 때 한 가지 유의해야 할 점은 도수가 지나치게 작은 곳들이 많으면 검증 결과가 부정확해진다는 것이다. 위 자료에서 '아마 사지 않을 것이다', '절대로 사지 않겠다', '모르겠다'에 대한 도수는 직업 수준을 막론하고 매우 낮다. 이같은 경우에는 아주 낮은 도수를 보이는 범주들을 병합할 필요가 있다. 다음은 범주를 병합하여 수정된 자료이다.

구 매 의 사	직업			합계
	사무직	육체 노동자	기타	
반드시 사겠다	87	54	22	163
아마 살 것이다	57	22	15	94
살 수도 안 살 수도 있다	19	8	4	31
기타	6	2	1	9
합계	169	86	42	297

이 4×3 분할표에서 계산된 χ^2 검증 통계량의 값은 3.43이고 유의확률은 0.75로서 통상의 유의수준 0.05보다 아주 크다. 따라서 구매 의사와 직업 수준 간에는 아무 연관성도 없다고 할 수 있다.

The SAS System

TABLE OF DECISION BY STATUS

DECISION STATUS

Frequency Percent Row Pct Col Pct	Blue	Else	White	Total
Else	2 0.67 22.22 2.33	1 0.34 11.11 2.38	6 2.02 66.67 3.55	9 3.03
Maybe	22 7.41 23.40 25.58	15 5.05 15.96 35.71	57 19.19 60.64 33.73	94 31.65
Must	54 18.18 33.13 62.79	22 7.41 13.50 52.38	87 29.29 53.37 51.48	163 54.88
Yes/No	8 2.69 25.81 9.30	4 1.35 12.90 9.52	19 6.40 61.29 11.24	31 10.44
Total	86 28.96	42 14.14	169 56.90	297 100.00

STATISTICS FOR TABLE OF DECISION BY STATUS

	Statistic	DF	Value	Prob	
==>	Chi-Square	6	3.430	0.753	<==
==>	Likelihood Ratio Chi-Square	6	3.463	0.749	<==
	Mantel-Haenszel Chi-Square	1	0.806	0.369	
	Phi Coefficient		0.107		
	Contingency Coefficient		0.107		
	Cramer's V		0.076		

Sample Size = 297
WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

3.5 독립성 검증을 하는 χ^2 통계량

- 카이 제곱(chi-square)
- 우도비 카이 제곱(likelihood ratio chi-square)
- 연속성 수정된 카이 제곱(continuity adjusted chi-square)

독립성 검증을 하는 χ^2 통계량들 중 'Chi-square'란 레이블이 붙은 통계량이 바로 앞서 소개한 가장 흔히 사용되는 피어슨(Pearson)의 χ^2 통계량이다. 나머지 두 개도 분할표에서 독립성 검증을 하는 χ^2 통계량들로, 고안된 배경은 다르지만 궁극적으로 카이제곱 분포를 따르는 것은 똑같다.

이들 중 연속성 수정이 된 카이제곱 통계량은 피어슨 카이제곱 통계량을 다소 보정한 것으로, 자료가 작을 때는 피어슨 카이제곱 통계량보다 연속성 수정된 카이제곱 통계량의 값을 읽는 편이 더 낫다. 그러나, 자료의 크면 두 값의 차이는 거의 없다. 연속성 수정된 카이 제곱 통계량의 값은 2×2 분할표에서만 출력된다.

어느 통계량이든 출력에서 Prob 밑의 값들이 대응되는 유의확률이니, 이 유의확률값이 0.05보다 작으면 두 분류변수가 독립이 아니라할 수 있으며, 유의확률값이 0에 아주 가까울수록 두 분류변수 간에는 독립성이 없다는 증거가 뚜렷해진다.

3.6 기타

때때로 다음과 같은 메시지가 출력되는 수가 있다.

> WARNING: 75% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

이는 전체 셀에서 기대도수가 5보다 작은 셀들이 75%에 이르기 때문에 카이제곱 통계량들이 독립성 검증에 적법하지 않을 수도 있다는 경고문이다. 독립성 검증을 할 때는 언제나 전체 관측수가 많고, 더불어 각 셀의 도수가 크면 클수록 좋다. 특히, 도수가 0인 셀들의 갯수가 적을수록 검증이 잘 된다.

그러나, 실제 데이타를 수집해서 분할표를 만들어보면 도수가 낮은 셀들이 나타나는 일이 많다. 이러한 경우 SAS에서는 내장된 기준에 의거하여 위와 같은 경고문을 출력시킨다. 도수의 높고 낮음을 가리는 기준값이라든가 또 도대체 낮은 도수의 셀들이 몇%일 때 정상적인 카이 제곱 검증이 되지 않는지에 대하여는 학자들 간에 이론이 많다. 여하튼 이런 경우에 마음이 불안하다면 Fisher의 정확 검증 통계량의 값을 읽어 독립성 유무를 판정하기를 권한다.

3.7 독립성 검증 사례

다음은 이화여대생 대상의 설문조사중 일부이다. 약 15200 명의 전체 학부생들 모집단에 대하여, 단과대학과 학년을 층으로 놓고 2단 층화랜덤추출기법을 사용하여 400여명의 랜덤 표본을 추출하여 설문지 조사를 실시하였다.

1. 호감이 가는 상대를 만났을 때 어떻게 하겠습니까?

① 좋아하는 감정을 드러내지 않는다.
② 상대에게 자신의 감정을 알게한다.

8. 애인의 여자 친구를 인정하겠습니까?

① 인정한다.
② 인정하지 않는다.

9. 애인관계에서 신체적 접촉을 어느 정도까지 허락하겠습니까?

① 전혀 허락하지 않겠다.
② 손, 팔짱, 어깨동무 정도
③ 입맞춤
④ 성관계만 빼고 허락한다.
⑤ 다 허락한다.

10. 애인이 있을 때 더 좋은 상대가 나타나면 어떻게 하겠습니까?

① 흔들리지 않겠다.
② 더 좋은 상대를 선택한다.
③ 둘 다 만난다.

다음과 같은 분석을 고려한다.

1) 8번 문항의 응답 패턴이 학년별로 차이가 있는지를 알고 싶다.
2) 1, 2학년을 저학년, 3, 4학년을 고학년이라 할 때 저학년과 고학년 간에 8번 문항의 응답 패턴의 차이가 있는지 알고 싶다.
3) 1번 문항에 ①이라 답한 학생들에 대하여 8번 문항의 응답 패턴이 학년별로 차이가 있는지를 알고 싶다.
4) 1번 문항에 2이라 답한 학생들에 대하여 8번 문항의 응답 패턴이 학년별로 차이가 있는지를 알고 싶다.

출력을 보면 전체적으로 인정과 불인정이 7:3으로 상당히 개방적임을 알 수 있다. 카이제곱 통계량에 대한 유의확률은 0.012로 학년간 차이가 뚜렷하다. 각 셀의 값들을 살펴보면 쉽게 알 수 있지만 주요 원인은 4학년의 응답 패턴이 다른 학년과 다르기 때문이다.

Crosstable: Question #8 for Grade

TABLE OF GRADE BY Q8

GRADE Q8

Frequency / Percent / Row Pct / Col Pct	1	2	Total
1	81	22	103
	20.25	5.50	25.75
	78.64	21.36	
	28.93	18.33	
2	73	30	103
	18.25	7.50	25.75
	70.87	29.13	
	26.07	25.00	
3	72	28	100
	18.00	7.00	25.00
	72.00	28.00	
	25.71	23.33	
4	54	40	94
	13.50	10.00	23.50
	57.45	42.55	
	19.29	33.33	
Total	280	120	400
	70.00	30.00	100.00

STATISTICS FOR TABLE OF GRADE BY Q8

Statistic	DF	Value	Prob
Chi-Square	3	10.944	0.012
Likelihood Ratio Chi-Square	3	10.759	0.013
Mantel-Haenszel Chi-Square	1	8.955	0.003
Phi Coefficient		0.165	
Contingency Coefficient		0.163	
Cramer's V		0.165	

Sample Size = 400

저학년일수록 개방적이고 고학년일수록 비개방적임을 알 수 있다.

Crosstable: Question #8 for Status

TABLE OF STATUS BY Q8

STATUS Q8

Frequency Percent Row Pct Col Pct	1	2	Total
High	126 31.50 64.95 45.00	68 17.00 35.05 56.67	194 48.50
Low	154 38.50 74.76 55.00	52 13.00 25.24 43.33	206 51.50
Total	280 70.00	120 30.00	400 100.00

STATISTICS FOR TABLE OF STATUS BY Q8

Statistic	DF	Value	Prob
Chi-Square	1	4.577	0.032
Likelihood Ratio Chi-Square	1	4.584	0.032
Continuity Adj. Chi-Square	1	4.122	0.042
Mantel-Haenszel Chi-Square	1	4.566	0.033
Fisher's Exact Test (Left)			0.021
(Right)			0.988
(2-Tail)			0.038
Phi Coefficient		-0.107	
Contingency Coefficient		0.106	
Cramer's V		-0.107	

Sample Size = 400

1번 문항에서 ①번에 대답한 학생들은 일종의 내숭파들이다. 이들에 국한된 8번 문항의 응답 패턴은 첫번째 출력과 비슷하다.

```
Crosstable: Question #8 for Grade when Question #1 is 1

                TABLE OF GRADE BY Q8

        GRADE     Q8

        Frequency|
        Percent  |
        Row Pct  |
        Col Pct  |       1|       2|  Total
        ---------+--------+--------+
              1  |     66 |     14 |     80
                 |  25.78 |   5.47 |  31.25
                 |  82.50 |  17.50 |
                 |  34.92 |  20.90 |
        ---------+--------+--------+
              2  |     50 |     16 |     66
                 |  19.53 |   6.25 |  25.78
                 |  75.76 |  24.24 |
                 |  26.46 |  23.88 |
        ---------+--------+--------+
              3  |     40 |     15 |     55
                 |  15.63 |   5.86 |  21.48
                 |  72.73 |  27.27 |
                 |  21.16 |  22.39 |
        ---------+--------+--------+
              4  |     33 |     22 |     55
                 |  12.89 |   8.59 |  21.48
                 |  60.00 |  40.00 |
                 |  17.46 |  32.84 |
        ---------+--------+--------+
        Total         189       67     256
                    73.83    26.17  100.00

        STATISTICS FOR TABLE OF GRADE BY Q8

Statistic                     DF      Value        Prob
-------------------------------------------------------
Chi-Square                     3      8.718       0.033
Likelihood Ratio Chi-Square    3      8.530       0.036
Mantel-Haenszel Chi-Square     1      8.168       0.004
Phi Coefficient                       0.185
Contingency Coefficient               0.181
Cramer's V                            0.185

Sample Size = 256
```

그러나, 아래에서 볼 수 있듯이, 1번 문항에서 ②번에 대답한 학생들에 대해서는 8번 문항의 응답 패턴이 학년과 무관하다.

결국 내숭파들이 졸업 때가 다가오면서 현실을 직시하게 되는 변모가 학년별 차이를 주도하고 있는 것이다. 그러나, 원래부터 개방적인 학생들은 내내 그렇다고 할 수 있다.

Crosstable: Question #8 for Grade when Question #1 is 2

TABLE OF GRADE BY Q8

GRADE	Q8 1	Q8 2	Total
1	15	8	23
	10.56	5.63	16.20
	65.22	34.78	
	16.48	15.69	
2	23	14	37
	16.20	9.86	26.06
	62.16	37.84	
	25.27	27.45	
3	32	12	44
	22.54	8.45	30.99
	72.73	27.27	
	35.16	23.53	
4	21	17	38
	14.79	11.97	26.76
	55.26	44.74	
	23.08	33.33	
Total	91	51	142
	64.08	35.92	100.00

STATISTICS FOR TABLE OF GRADE BY Q8

Statistic	DF	Value	Prob
Chi-Square	3	2.785	0.426
Likelihood Ratio Chi-Square	3	2.809	0.422
Mantel-Haenszel Chi-Square	1	0.282	0.595
Phi Coefficient		0.140	
Contingency Coefficient		0.139	
Cramer's V		0.140	

Sample Size = 142

4. 산점도, 상관분석과 회귀분석

산점도(scatter plot)와 상관분석(correlation analysis)는 두 변수 간의 연관도를 보는 방법이다. 회귀분석은 하나의 변수의 추이를 다른 여러 변수들로 설명하고자 할 때 사용하는 기법이다. 회귀분석은 가장 사용빈도가 높은 통계기법의 하나이다.

4.1 산점도

SAS에서 변수들 간의 산점도를 그릴 때는 PLOT 절차를 이용한다. 특히 고해상도 그래프가 필요할 때는 SAS/GRAPH의 GPLOT 절차를 이용하는 편이 더 낫다. 참고로 G3D 절차를 이용하면 삼차원도(3-dimensional plot)를 그릴 수도 있다. PLOT 절차에서 산점도를 그릴 때는 SAS 자료에 포함된 두 변수의 값들을 입력하여 대응되는 한 쌍의 값들이 교차되는 지점을 점으로 나타내게 된다.

4.2 상관분석

상관분석(correlation analysis)은 변수들 간의 관련성을 연구하는 가장 초보적인 분석법이다. 특히, 두 변수 간의 선형 연관성의 강도를 측정할 때는 흔히 **피어슨 상관계수**(Pearson correlation coefficient) r을 사용한다. SAS에서는 CORR 절차로 상관분석을 수행한다.

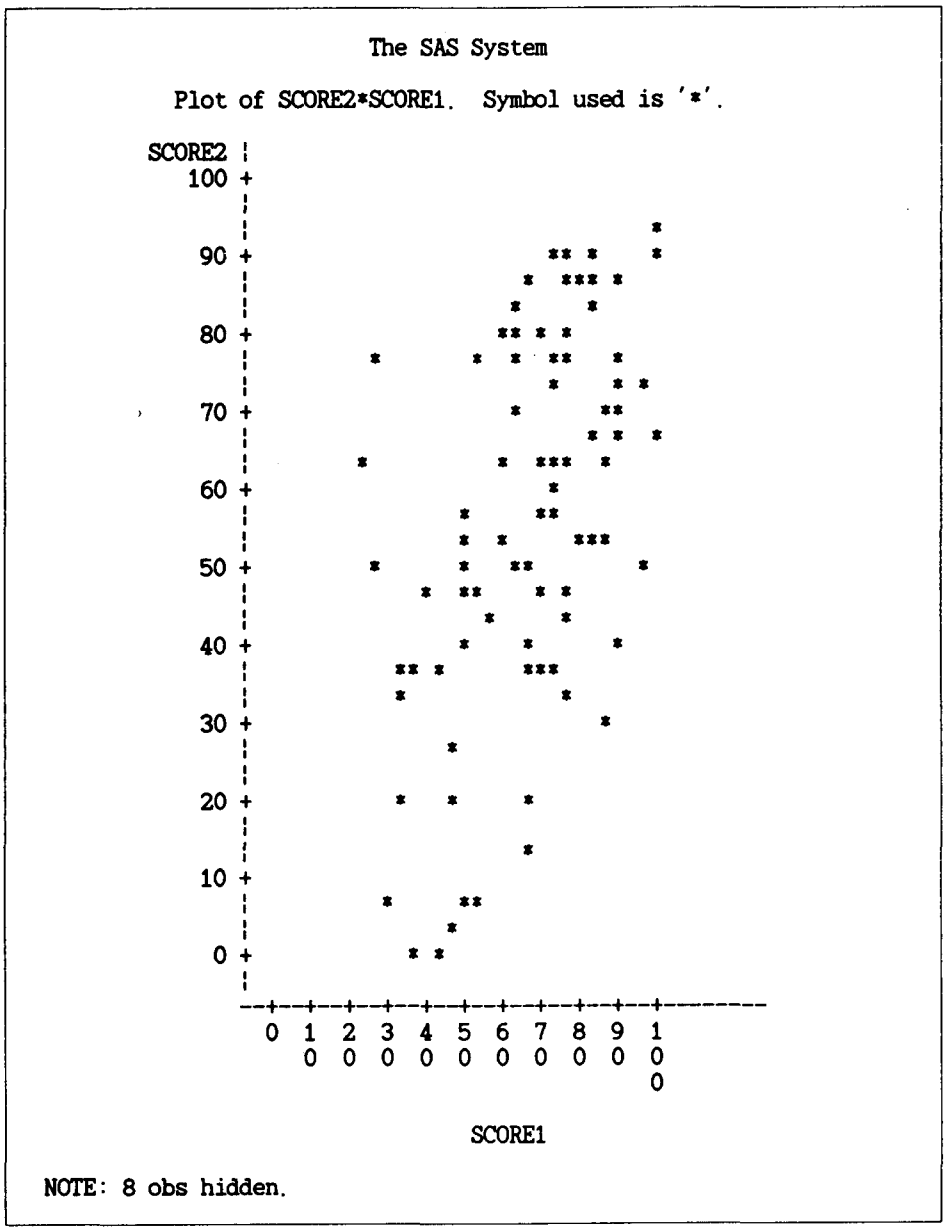
변수 X와 Y에 대한 n 쌍의 데이타를 $(X_1,Y_1),\cdots,(X_n,Y_n)$으로 표기할 때 X와 Y 간의 상관 계수 r의 공식은 다음과 같다.

$$r = \frac{\sum_i^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_i^n (X_i - \overline{X})^2 \sum_i^n (Y_i - \overline{Y})^2}}$$

r의 값은 언제나 -1에서 $+1$ 사이에 있다. -1이면 두 변수 간의 관계가 완벽한 역선형 관계, 즉, 하나가 감소하면 다른 하나는 증가하고 또 하나가 증가하면 다른 하나는 감소하는 관계이다. 반대로 $+1$이면 완벽한 정선형 관계로서 두 변수는 같이 증감한다. 상관 계수의 값이 0이면 두 변수 간의 선형 관계는 없으며, 다시 말해서 두 변수 간에는 아무 관계도 없다.

따라서, 상관 계수가 0에 가까우면 두 변수 간에는 서로를 설명할 근거가 없다고 할 수 있다. 그러나, 상관 계수의 값이 어느 정도라야 서로 무관하다고 할 수 있는지는 쉽게 말할 수

없다. 사회 경제 분야에서는 상관 계수가 ±0.5를 넘는 경우가 드물기 때문에 ±0.3 정도라도 분석을 계속 진행한다. 그러나 ±0.3은 정교한 화학 실험 등에서는 분석할 가치가 없는 값이다.



상관도가 높더라도 변수 간의 인과성을 쉽게 말할 수는 없다. 즉, 두 변수 간의 연관성이 높더라도 어느 한 변수가 다른 변수의 원인이라는 말은 못한다.

통계 패키지를 이용하면 상관계수는 쉽게 얻을 수 있는데, 상담을 하다보면 컴퓨터로 출력된 상관계수의 해석에 오류를 범하는 경우를 자주 접한다. 예를 들어, 대다수의 논문이나 보고서에는 삽입된 상관 계수에는 0.78**와 같이 통계적 유의도를 나타내는 별표를 덧붙이는 것을 자주 볼 수 있다. 그러나, 이때 검증한 것은 두 변수 간의 모집단 상관계수가 0인지 아닌지에 대한 것으로서 이 유의도의 실용성은 거의 없다. 특히 검증 방법상에 내재된 문제점 때문에 표본 크기 n이 크면 항상 상관계수는 유의하다고 나온다. 만일 0.02**라고 상관계수가 나온다면 두 변수 간의 연관도를 있다고 해야할까 없다고 해야할까? '**'는 두 변수 간의 상관계수가 통계적으로 0이 아니라는 의미이이만, 실제 계산된 상관계수 0.02는 0에 아주 가까우므로 두 변수 간의 상관계수를 0이라 간주하여도 무리가 없어 보인다. 이런 경우에는 물론 후자의 해석을 따라야 하고 유의도는 무시한다. 결론적으로 거의 대부분의 경우 상관계수의 유의도는 실용적 가치가 없다.

[예] 뒤따르는 표는 다섯 변수에 대한 33 개의 측정값들이다. 이 자료에 대한 상관계수를 계산하면 다음과 같다. X_1과 X_5 간의 상관 계수가 −0.73으로 가장 크며 다른 계수들에 비해서 두드러지게 크다.

	X_1	X_2	X_3	X_4	X_5
X_1	1.00000	0.32872	0.16767	0.05191	−0.73081
X_2	0.32872	1.00000	−0.14550	0.18033	−0.21204
X_3	0.16767	−0.14550	1.00000	0.24134	−0.05541
X_4	0.05191	0.18033	0.24134	1.00000	0.31267
X_5	−0.73081	−0.21204	−0.05541	0.31267	1.00000

4.3 회귀분석

회귀분석(regression analysis)은 여러 개의 독립변수들로 하나의 종속변수를 설명하는 함수식을 만들고자 할 때 사용한다. SAS에서 회귀분석은 REG 절차에서 담당한다. 회귀모형으로는 주로 일차 함수식이 사용되며, 그렇기 때문에 **선형회귀분석**(linear regression analysis)이 주종을 이룬다. 종속변수를 Y, 독립변수들을 $X_1, \cdots, X_{k-1}$이라 하자. 즉, $(k-1)$개의 독립변수들로 종속변수를 설명하는 것이다. 이때 선형회귀모형은 다음과 같다. 여기서 추정해야할 모수(parameter)의 갯수는 k 개이다.

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$$

X_1(℃)	X_2(cm)	X_3(mm)	X_4(min)	X_5(ml)
6	9.9	5.7	1.6	2.12
1	9.3	6.4	3.0	3.39
-2	9.4	5.7	3.4	3.61
11	9.1	6.1	3.4	1.72
-1	6.9	6.0	3.0	1.80
2	9.3	5.7	4.4	3.21
5	7.9	5.9	2.2	2.59
1	7.4	6.2	2.2	3.25
1	7.3	5.5	1.9	2.86
3	8.8	5.2	0.2	2.32
11	9.8	5.7	4.2	1.57
9	10.5	6.1	2.4	1.50
5	9.1	6.4	3.4	2.69
-3	10.1	5.5	3.0	4.06
1	7.2	5.5	0.2	1.98
8	11.7	6.0	3.9	2.29
-2	8.7	5.5	2.2	3.55
3	7.6	6.2	4.4	3.31
6	8.6	5.9	0.2	1.83
10	10.9	5.6	2.4	1.69
4	7.6	5.8	2.4	2.42
5	7.3	5.8	4.4	2.98
5	9.2	5.2	1.6	1.84
3	7.0	6.0	1.9	2.48
8	7.2	5.5	1.6	2.83
8	7.0	6.4	4.1	2.41
6	8.8	6.2	1.9	1.78
6	10.1	5.4	2.2	2.22
3	12.1	5.4	4.1	2.72
5	7.7	6.2	1.6	2.36
1	7.8	6.8	2.4	2.81
8	11.5	6.2	1.9	1.64
10	10.4	6.4	2.2	1.82

회귀분석을 시도할 때는 우선적으로 회귀모형의 유의성을 검증해야 한다. 회귀모형이 유의하다는 말은 가정된 회귀모형이 데이타를 잘 설명하고 있음을 뜻한다. 회귀모형의 유의성은 분산분석으로 검증한다. 이 경우의 귀무가설은 '회귀모형이 유의하지 않다'이기 때문에 귀무가설을 기각해야 분석을 계속 진행할 수 있다. 회귀모형의 유의성이 검증되면 **회귀계수** (regression coefficients) β들을 추정해서 적합된 회귀식을 정립해야 한다. 그러나, 이들 계산은 전적으로 행렬대수에 의존하기 때문에 손으로 풀 수는 없다.

여기서는 제시된 자료를 이용하여 기초적인 회귀분석의 방법과 해석법을 소개한다. 자료에서 종속변수를 X_5로 잡고 나머지 네 개의 변수들을 독립변수로 간주하자. 이 경우 회귀모형은 다음과 같다.

$$X_5 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

뒤따르는 출력의 상단에 나와있는 분산분석표에서 유의확률 p 값(출력에서 Prob>F 밑의

값)이 0.0001이므로 통계적으로 극히 유의한 회귀 관계가 있다. 결정계수(coefficient of determination) R^2의 값이 0.6589이므로 회귀모형이 자료 변동의 약 66 %를 설명하고 있다. 적합된 회귀선의 식은 다음과 같다.

$$\widehat{X_5} = 2.958 - 0.129X_1 - 0.019X_2 - 0.0462X_3 + 0.209X_4$$

여기서 X_5 위의 고깔모자(^)는 추정되었음을 표시하는 부호이다.

The SAS System

Dependent Variable: X5

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	4	9.71735	2.42934	13.524	0.0001
Error	28	5.02986	0.17964		
C Total	32	14.74721			

R-square 0.6589

Parameter Estimates

Variable	DF	Parameter Estimate	T for H0: Parameter=0	Prob > |T|	Standardized Estimate
INTERCEP	1	2.958283	2.169	0.0387	0.00000000
X1	1	-0.129319	-6.075	0.0001	-0.73174667
X2	1	-0.018785	-0.334	0.7410	-0.04110811
X3	1	-0.046215	-0.223	0.8252	-0.02666415
X4	1	0.208755	3.114	0.0042	0.36450078

회귀분석류의 통계 방법들에서는 **변수선택**(variable selection)이 중요한 분석 과제 중의 하나이다. 독립 변수가 많은 모형은 피해야 한다. 가능한 한 적은 수의 독립 변수로 종속 변수를 잘 설명하는 모형을 찾아야 한다.

변수선택을 할 때는 우선 각 회귀계수들의 통계적 유의도를 살펴보아야 한다. 회귀계수의 유의도란 대응되는 독립변수가 모형에 반드시 들어갈 필요가 있는지를 검증하는 것으로, 유의성이 없다면 그 독립변수를 모형에서 제거하는 것을 고려해야 한다. 출력에서 회귀계수의 유의도에 대한 p 값(출력에서 Prob>|T| 밑의 값)을 살펴보면 X_1과 X_4는 통계적으로 극히 유의하며, X_2와 X_3는 유의확률이 각각 0.741과 0.825로서 전혀 유의하지 않다. 따라서 X_2와 X_3는 종속변수를 설명하는데 큰 도움이 되지 못한다. 이때 모형의 절편(intercept)인 β_0에

대해서는 대부분의 경우 검증하지 않는다.

회귀분석에서 적합된 모형을 해석할 때 많은 사람들이 단순히 회귀계수들의 값을 비교하여 독립변수의 상대적인 중요도를 판정하는데 이는 전적으로 잘못된 것이다. 회귀계수는 독립변수의 단위를 바꾸면 그 배수만큼 달라지게 되어있다. 즉, 회귀계수는 단위에 의존한다. 예를 들어, 동일한 자료인데 한 사람은 *cm* 단위의 독립변수를 사용하고 다른 사람은 *mm* 단위의 독립변수를 사용했다면 두 사람의 회귀계수에는 10 배의 차이가 난다. 따라서, 단위에 의존하는 회귀계수들의 값을 기초로 대응되는 독립변수의 중요도를 판정할 수 없다. 즉, 출력에서 β_1은 −0.129로 추정되었고 β_4는 +0.209로 추정되었다. 부호를 무시하면 β_4는 β_1의 약 2 배이므로 X_4가 X_1보다 2 배 중요하다고 판정한다면 이는 오해이다.

회귀계수의 상대적 중요도를 판정할 때는 항상 **표준화 회귀계수**(standardized regression coefficients)의 값을 기초로 해야 한다. 표준화 회귀계수의 값은 출력의 아래 오른쪽에 위치한 'Standardized Estimate'의 밑에 출력되어 있다. 절편에 대한 표준화 회귀계수의 값은 항상 0이다. 출력에서 독립변수들에 대한 표준화 회귀계수들은 다음과 같다.

독립 변수	표준화 회귀계수
X_1	−0.732
X_2	−0.041
X_3	−0.027
X_4	+0.365

여기서도 확인할 수 있지만 X_2와 X_3에 대한 표준화 회귀계수의 값은 0에 가까우므로 종속변수를 설명하는데 별 도움이 되지 못한다. 그리고 부호를 무시하면 X_1의 표준화 회귀계수가 X_4의 표준화 회귀계수의 약 2 배이므로 X_1이 X_4보다 상대적으로 두 배 가량 중요하다고 하겠다. 원래의 회귀계수들로 해석할 때와는 정반대 결과임을 유의하라.

그러나, 표준화 회귀계수로 독립변수들 간의 상대적 중요도를 비교할 때도 몇 가지 제한이 있는바, 그중 가장 중요한 것이 독립변수들 간에 상관도가 높지 않아야 한다는 점이다. 상관도가 높으면 독립변수들끼리 서로 영향을 미쳐 표준화 회귀계수의 값들이 왜곡되기 때문이다. 4.2 절에서 자료에 대한 상관계수들을 예시하였는데 이를 다시 보면 독립변수들 간에 상관도가 크다고 할 것이 없으므로 이 경우에는 표준화 회귀계수로 상대적 중요도를 사정해도 문제될 것이 없다.

여하튼 이상의 해석에서 X_2와 X_3는 모형에서 제거할 수 있다는 확신을 얻을 수 있다. 뒤따르는 출력은 X_2와 X_3를 모형에서 제거하고 X_1과 X_4만으로 종속변수를 적합하는 경우에 대한 출력 결과이다. 즉, 이때의 모형은 다음과 같다.

$$X_5 = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + \varepsilon$$

적합된 회귀선의 식은 다음과 같다.

$$\widehat{X_5} = 2.552 - 0.132X_1 + 0.201X_4$$

원래의 4 개의 독립변수 모두를 사용해서 적합했을 때와 비교하여 회귀계수들 간에 별 차이가 없다. 이는 독립변수들 간에 서로 상관관계가 별로 없었음을 뒷받침한다. 독립 변수들 간의 상관 관계가 크면 **다공선성**(multicollinearity)이 존재한다고 하는데, 우리의 자료에서는 다공선성이 없다.

The SAS System

Dependent Variable: X5

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	9.69388	4.84694	28.775	0.0001
Error	30	5.05332	0.16844		
C Total	32	14.74721			

R-square 0.6573

Parameter Estimates

Variable	DF	Parameter Estimate	T for H0: Parameter=0	Prob > |T|	Standardized Estimate
INTERCEP	1	2.552037	13.618	0.0001	0.00000000
X1	1	-0.132378	-6.999	0.0001	-0.74905829
X4	1	0.201339	3.285	0.0026	0.35155148

출력에서 결정계수의 값은 0.6573으로 앞의 결정계수 0.6589에 비하여 불과 0.0016만큼 감소하였다. 이는 곧 X_2와 X_3를 모형에서 제거하여도 자료의 변화 패턴에 대한 설명력의 감소가 0.16%에 불과하다는 의미이다.

회귀모형에 독립변수를 추가하면 할수록 결정계수의 값은 점점 증가한다. 심지어 종속변수를 설명하는데 얼토당토않은 독립변수를 추가하여도 결정계수의 값은 증가한다. 그렇기 때문에 어떤 독립변수의 가감에 따라 결정계수의 값의 차이가 미미하면 이 독립변수가 종속변수를 설명하는 정도가 약하다는 뜻이며 따라서 이 독립변수를 모형에서 제거하여 전체 모형을 단순화시킬 수 있다.

여기서 표준화 회귀계수들의 값은 −0.749와 +0.352로 앞의 출력과 비교하여 별 차이가 없으며, 이들에 대한 해석 역시 전과 마찬가지이다.

회귀분석을 할 때는 언제나 종속변수와 독립변수들 간의 상관계수를 확인하여야 한다. 앞서 제시한 상관계수를 살펴보면 X_1과 X_5 간의 상관계수가 −0.731로 가장 높고 X_4와 X_5 간의 상관계수가 +0.313으로 그 다음이다. 그리고, 독립 변수들간의 상관도는 거의 없다. 이런 경우에는 회귀분석을 하기 전이라도 X_5를 설명하는데 X_1과 X_4가 가장 중요하리라고 미리 짐작할 수 있다.

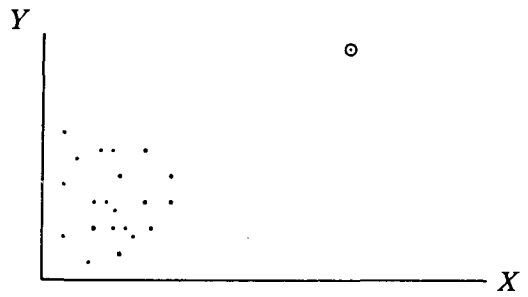
종속변수와 독립변수 간의 상관도가 높으면 높을수록 회귀분석은 잘 된다. 그러나, 독립변수들 간의 상관도가 높으면 회귀·적합의 결과가 잘 나오든 잘 나오지 않든 해석상 문제가 많게 된다.

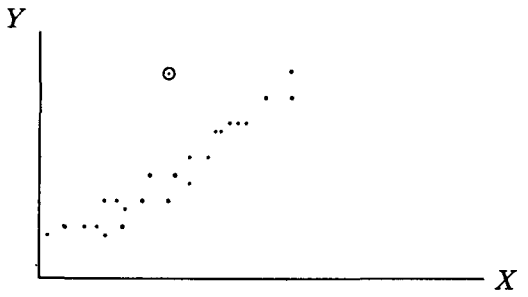
4.4 영향진단

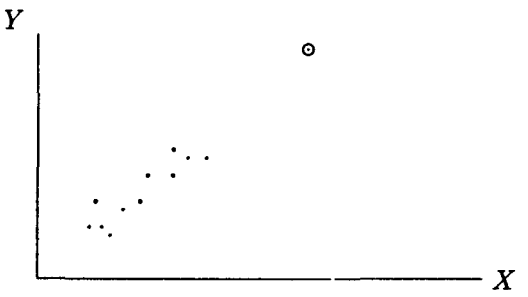
일단 초기 회귀분석을 한 후에는 반드시 **잔차분석**(residual analysis)과 **영향진단**(influence diagnostics)이 뒤따라야 한다. 잔차분석으로는 모형이 올바른지 판정하며, 영향진단에서는 데이타 중 이상점이 있는지 여부를 가리게 되며, 이상점이더라도 회귀분석에 도움이 되는지 해를 끼치는지 판별한다. 이외에도 **공선성진단**(collinearity diagnostics)도 해야하지만 이 부분에 대한 간단한 설명은 어렵다.

[mlr-1]부터 [mlr-5]까지의 출력을 보자. Residual은 각 관측에 대한 잔차이며, Student Residual은 일종의 표준화된 잔차로, 정확히는 **스튜던트화 잔차**(Studentized residual)이라 부른다. Student Residual 옆의 그림은 스튜던트화 잔차들의 상대적 위치를 표시한 것으로, 별표가 3개 이상 찍힌 관측들은 회귀선 적합에 영향을 주는 이상점일 가능성이 있다. 이런 관측들에 대하여는 좀 더 세밀한 분석이 필요하다.

영향진단은 비교적 최근에 개발된 회귀진단 기법이다. 영향분석은 회귀분석에 가장 큰 영향을 주는 관측(들)을 찾는 기법이다. 여기서 한 가지 유념해야할 사실은, 분석에 악영향을 미치는 이상점도 영향이 큰 관측에 들어가나, 영향이 큰 관측이 반드시 나쁘지는 않다는 점이다. 영향이 큰 관측(influential observation)이란, 그 점을 집어넣고 회귀분석을 할 때와 빼고서 회귀분석을 할 때 회귀계수값이라든가 추측값에 커다란 변동이 생기는 관측을 말한다. 다음의 세 그림에서 고립된 점들은 둘 다 영향이 큰 관측들이다.







첫번째 그림에서 회귀선의 기울기는 전적으로 관측 ⊙에 의존함을 쉽게 알 수 있을 것이다. 즉, 관측 ⊙는 아주 영향이 큰(highly influential) 관측이라 할 수 있다. 그러나, 관측 ⊙를 이상점으로 판정할 근거는 전혀 없다. 반면에 두번째 그림에서 관측 ⊙가 기울기 추정에 미치는 영향은 첫번째 그림에 비하여 미미하나, 절편 추정에는 아주 영향이 크리라 판단되며, 나머지 관측들이 보여주는 경향에서 벗어나 있으므로 영향이 큰 관측인 동시에 이상점이다. 세번째 그림에서도 관측 ⊙는 영향이 큰 관측인데, 이 경우에는 회귀선 적합에 큰 도움을 주는 관측이라 할 수 있다. 즉, 관측 ⊙는 나머지 관측들의 경향을 그대로 따르고 있으므로, 관측 ⊙가 있으나 없으나 회귀계수들의 추정에는 큰 영향을 미치지 못하나, 이 관측 덕분에 추정된 회귀계수들의 오차가 무척 줄어들 것이라 예측할 수 있다.

회귀선의 적합에 영향이 큰 관측을 탐색하는데 큰 역할을 하는 도구는 출력에서 Hat Diag H 아래의 값들로서 **모자대각원소**(hat diagonal elements)라 부른다. 모자대각원소 h_i는 다

음과 같이 언제나 1 이하이며 $1/n$ 이상의 값을 갖는다.

$$1/n \le h_i \le 1$$

h_i는 i번째 관측이 자료의 중심에서 얼마나 멀리 떨어져있나를 재는 측도로서, 일종의 표준화된 거리라고 생각하면 된다. 그리고, 모자대각원소 h_i를 지레(leverage)라 하며, 특히 h_i의 값이 1에 가까운 관측, 즉, 자료의 중심에서 멀리 떨어져있는 관측을 큰 지레점(high leverage point)이라 한다. 우리가 무거운 바위를 들어올리기 위하여 지렛대를 이용할 때, 받침대를 바위 가깝게 놓고 긴 막대를 사용하면 큰 힘을 쓰지 않고도 쉽게 바위를 들어올릴 수 있음을 누구나 알 것이다. 회귀계수 추정에 강한 영향력을 발휘하는 관측은 따라서 자료의 중심에서 멀리 떨어져있는 관측이며, 이런 관측이 큰 지레점에 해당한다. k가 모형에 가정한 회귀계수들의 숫자일 때, **h_i의 값이 대체로 $(2k/n)$보다 큰 관측은 영향이 큰 관측이라고 판단한다.**

출력에서 Rstudent는 앞에서 소개한 스튜던트화 잔차와는 조금 다른 종류의 표준화된 잔차로서, **외부 스튜던트화 잔차**(externally Studentized residual)한다. 이 통계량은 이상점을 탐지하는데 쓰이며, 그렇기 때문에 이상점 진단 통계량이라고도 불리운다. 지레는 X축 방향으로 지나친 점을 탐지하는 반면, R-Student는 Y축 방향으로 지나친 점을 탐지하는데 쓰인다. 대체로 **R-Student의 절대값이 3보다 크면 이상점의 용의가 있다**고 판단한다. 출력에서 5번, 25번 관측이 여기에 해당한다.

출력에서 Dffits는 하나의 관측이 추측값 y_i에 미치는 영향을 재는 측도로서, 일반적으로 DFFITS(디핏츠)로 표기한다. **디핏츠의 절대값이 2보다 큰 관측은 추측값에 영향을 미치는 관측이라 판정**한다.

출력을 보면, 각 회귀변수에 대한 Dfbetas 값들이 나와있다. 일반적인 표기는 DFBETAS이며, '디프베이타즈'라 읽는다. **어떤 회귀변수에 대한 디프베이타즈의 절대값들 중 특히 2보다 큰 값이 있으면, 대응되는 관측이 그 회귀 계수 추정에 영향을 크게 미친다**고 판정한다.

디프베이타즈는 개별적인 회귀계수 추정에 영향을 미치는 관측을 찾는 측도인 반면, 쿡의 D (Cook's D)는 추정하려는 회귀계수들 전반에 영향을 미치는 관측들을 찾는 복합 측도이다. 이 값이 아주 큰 관측들은 회귀계수 - 특별히 어느 회귀계수인지는 모르나 - 추정에 영향이 크다고 판정한다. 출력에서 5번과 아마도 25번 관측이 여기에 해당한다.

Cov Ratio는 임의의 관측이 회귀선 적합을 향상시키는 지 여부를 가리는 데 사용하는 통계량으로, 원래의 표기는 COVRATIO이며, '코브레이쇼'라 읽는다. **COVRATIO의 값이 1보다 큰 관측들은 회귀선 적합에 도움이 되는 점들이며, 1보다 작은 관측들은 회귀선 적합에 해가 되는 관측들이다.** 특히, 범위 $(1-3k/n,\ 1+3k/n)$을 벗어나는 관측들에 주의를 기울여야 한다. **코브레이쇼의 값이 $1+(3k/n)$보다 큰 관측은 회귀 분석을 아주 향상시키는 관측이며, 반대로 $1-(3k/n)$보다 작은 관측은 회귀 분석에 도움이 못되는 관측이다.** 출

력에서 5번, 25번 관측들은 회귀 분석을 향상시키지 못하는 관측들이다.

[mlrplus] 출력들은 이상의 분석을 토대로 5번과 25번 관측을 제외한 나머지 관측들로만 다시 회귀분석을 하는 경우에 대한 것이다.

독립변수로 X1과 X4만 사용한 경우의 출력을 보면 5번과 25번 관측을 제거하기 전에 비하여 R^2 값이 0.6573에서 0.9020으로 현저한 향상을 보이고 있다.

[mlrplus-4]를 보면 14번 관측이 새로운 이상점일 가능성이 있지만 현재 가정된 회귀모형이 자료를 아주 잘 설명하고 있기 때문에 더 이상의 영향진단은 불필요하다.

[mlr-1] The SAS System

Correlation Analysis

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 33

	X1	X2	X3	X4	X5
X1	1.00000 0.0	0.32872 0.0618	0.16767 0.3510	0.05191 0.7742	-0.73081 0.0001
X2	0.32872 0.0618	1.00000 0.0	-0.14550 0.4191	0.18033 0.3153	-0.21204 0.2362
X3	0.16767 0.3510	-0.14550 0.4191	1.00000 0.0	0.24134 0.1760	-0.05541 0.7594
X4	0.05191 0.7742	0.18033 0.3153	0.24134 0.1760	1.00000 0.0	0.31267 0.0765
X5	-0.73081 0.0001	-0.21204 0.2362	-0.05541 0.7594	0.31267 0.0765	1.00000 0.0

[mlr-2] The SAS System

Dependent Variable: X5 < 독립변수가 4개인 경우 >

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	4	9.71735	2.42934	13.524	0.0001
Error	28	5.02986	0.17964		
C Total	32	14.74721			

Root MSE 0.42384 R-square 0.6589

Parameter Estimates

Variable	DF	Parameter Estimate	T for H0: Parameter=0	Prob > |T|	Standardized Estimate
INTERCEP	1	2.958283	2.169	0.0387	0.00000000
X1	1	-0.129319	-6.075	0.0001	-0.73174667
X2	1	-0.018785	-0.334	0.7410	-0.04110811
X3	1	-0.046215	-0.223	0.8252	-0.02666415
X4	1	0.208755	3.114	0.0042	0.36450078

Dependent Variable: X5 < 독립변수가 2개인 경우 >

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	9.69388	4.84694	28.775	0.0001
Error	30	5.05332	0.16844		
C Total	32	14.74721			

Root MSE 0.41042 R-square 0.6573

Parameter Estimates

Variable	DF	Parameter Estimate	T for H0: Parameter=0	Prob > |T|	Standardized Estimate
INTERCEP	1	2.552037	13.618	0.0001	0.00000000
X1	1	-0.132378	-6.999	0.0001	-0.74905829
X4	1	0.201339	3.285	0.0026	0.35155148

[mlr-3] The SAS System

Obs	Dep Var X5	Predict Value	Std Err Predict	Lower95% Mean	Upper95% Mean	Lower95% Predict	Upper95% Predict
1	2.1200	2.0799	0.097	1.8812	2.2787	1.2185	2.9413
2	3.3900	3.0237	0.102	2.8159	3.2314	2.1601	3.8872
3	3.6100	3.5013	0.153	3.1887	3.8140	2.6067	4.3959
4	1.7200	1.7804	0.150	1.4740	2.0869	0.8880	2.6729
5	1.8000	3.2884	0.130	3.0235	3.5534	2.4094	4.1675
6	3.2100	3.1732	0.144	2.8788	3.4675	2.2848	4.0615
7	2.5900	2.3331	0.075	2.1793	2.4869	1.4809	3.1853
8	3.2500	2.8626	0.098	2.6618	3.0634	2.0007	3.7245
9	2.8600	2.8022	0.103	2.5914	3.0130	1.9379	3.6665
10	2.3200	2.1952	0.161	1.8655	2.5249	1.2945	3.0959
11	1.5700	1.9415	0.172	1.5911	2.2919	1.0330	2.8500
12	1.5000	1.8438	0.112	1.6142	2.0735	0.9748	2.7129
13	2.6900	2.5747	0.089	2.3930	2.7564	1.7170	3.4324
14	4.0600	3.5532	0.162	3.2228	3.8836	2.6522	4.4542
15	1.9800	2.4599	0.170	2.1121	2.8078	1.5524	3.3674
16	2.2900	2.2782	0.126	2.0204	2.5361	1.4013	3.1552
17	3.5500	3.2597	0.142	2.9696	3.5499	2.3728	4.1467
18	3.3100	3.0408	0.138	2.7582	3.3234	2.1562	3.9253
19	1.8300	1.7980	0.164	1.4624	2.1337	0.8952	2.7009
20	1.6900	1.7115	0.128	1.4509	1.9720	0.8337	2.5892
21	2.4200	2.5057	0.072	2.3578	2.6537	1.6546	3.3569
22	2.9800	2.7760	0.134	2.5016	3.0504	1.8941	3.6580
23	1.8400	2.2123	0.093	2.0228	2.4018	1.3529	3.0716
24	2.4800	2.5374	0.085	2.3630	2.7119	1.6813	3.3936
25	2.8300	1.8152	0.115	1.5793	2.0510	0.9444	2.6859
26	2.4100	2.3185	0.134	2.0441	2.5929	1.4365	3.2005
27	1.7800	2.1403	0.087	1.9619	2.3187	1.2834	2.9973
28	2.2200	2.2007	0.080	2.0365	2.3649	1.3466	3.0548
29	2.7200	2.9804	0.123	2.7282	3.2326	2.1051	3.8557
30	2.3600	2.2123	0.093	2.0228	2.4018	1.3529	3.0716
31	2.8100	2.9029	0.097	2.7050	3.1008	2.0416	3.7641
32	1.6400	1.8756	0.107	1.6573	2.0939	1.0094	2.7417
33	1.8200	1.6712	0.130	1.4067	1.9357	0.7923	2.5501

[mlr-4] The SAS System

Obs	Residual	Std Err Residual	Student Residual	-2-1-0 1 2	Cook's D	Rstudent
1	0.0401	0.399	0.101		0.000	0.0989
2	0.3663	0.398	0.921	*	0.019	0.9189
3	0.1087	0.381	0.285		0.004	0.2809
4	-0.0604	0.382	-0.158		0.001	-0.1556
5	-1.4884	0.389	-3.823	******	0.541	-5.2479
6	0.0368	0.384	0.096		0.000	0.0942
7	0.2569	0.403	0.637	*	0.005	0.6304
8	0.3874	0.398	0.972	*	0.019	0.9713
9	0.0578	0.397	0.146		0.000	0.1431
10	0.1248	0.377	0.331		0.007	0.3259
11	-0.3715	0.373	-0.996	*	0.070	-0.9963
12	-0.3438	0.395	-0.871	*	0.021	-0.8675
13	0.1153	0.401	0.288		0.001	0.2833
14	0.5068	0.377	1.344	**	0.111	1.3627
15	-0.4799	0.373	-1.285	**	0.115	-1.3000
16	0.0118	0.391	0.030		0.000	0.0296
17	0.2903	0.385	0.754	*	0.026	0.7483
18	0.2692	0.386	0.697	*	0.021	0.6906
19	0.0320	0.376	0.085		0.000	0.0836
20	-0.0215	0.390	-0.055		0.000	-0.0541
21	-0.0857	0.404	-0.212		0.000	-0.2088
22	0.2040	0.388	0.526	*	0.011	0.5195
23	-0.3723	0.400	-0.931	*	0.016	-0.9291
24	-0.0574	0.401	-0.143		0.000	-0.1407
25	1.0148	0.394	2.577	*****	0.190	2.8711
26	0.0915	0.388	0.236		0.002	0.2322
27	-0.3603	0.401	-0.898	*	0.013	-0.8955
28	0.0193	0.402	0.048		0.000	0.0471
29	-0.2604	0.391	-0.665	*	0.015	-0.6590
30	0.1477	0.400	0.369		0.002	0.3641
31	-0.0929	0.399	-0.233		0.001	-0.2292
32	-0.2356	0.396	-0.594	*	0.009	-0.5879
33	0.1488	0.389	0.382		0.005	0.3766

[mlr-5] The SAS System

Obs	Hat Diag H	Cov Ratio	Dffits	INTERCEP Dfbetas	X1 Dfbetas	X4 Dfbetas
1	0.0562	1.1718	0.0241	0.0154	0.0080	-0.0147
2	0.0614	1.0822	0.2351	0.0719	-0.1544	0.0727
3	0.1391	1.2755	0.1129	0.0254	-0.0921	0.0434
4	0.1337	1.2747	-0.0611	0.0267	-0.0493	-0.0188
5	0.0999	0.1660	-1.7487	-0.6266	1.4100	-0.4502
6	0.1233	1.2616	0.0353	-0.0113	-0.0128	0.0285
7	0.0337	1.0998	0.1177	0.0625	0.0178	-0.0336
8	0.0574	1.0669	0.2397	0.1725	-0.1566	-0.0429
9	0.0633	1.1793	0.0372	0.0308	-0.0228	-0.0130
10	0.1547	1.2954	0.1394	0.1334	-0.0173	-0.1227
11	0.1748	1.2127	-0.4586	0.2811	-0.3167	-0.2543
12	0.0751	1.1084	-0.2472	0.0014	-0.1899	0.0290
13	0.0470	1.1521	0.0629	-0.0137	0.0054	0.0368
14	0.1554	1.0881	0.5846	0.2235	-0.5146	0.1278
15	0.1722	1.1284	-0.5930	-0.5912	0.2015	0.4880
16	0.0947	1.2227	0.0096	-0.0051	0.0048	0.0060
17	0.1198	1.1876	0.2761	0.1823	-0.2351	-0.0285
18	0.1137	1.1894	0.2473	-0.0958	-0.0597	0.2061
19	0.1603	1.3175	0.0365	0.0292	0.0082	-0.0322
20	0.0966	1.2251	-0.0177	0.0012	-0.0146	0.0020
21	0.0311	1.1375	-0.0374	-0.0195	0.0042	0.0043
22	0.1072	1.2060	0.1800	-0.0924	0.0059	0.1518
23	0.0511	1.0684	-0.2156	-0.1620	-0.0309	0.1355
24	0.0433	1.1548	-0.0300	-0.0246	0.0089	0.0133
25	0.0792	0.5676	0.8420	0.3402	0.5107	-0.4465
26	0.1072	1.2330	0.0804	-0.0462	0.0372	0.0551
27	0.0453	1.0685	-0.1950	-0.1054	-0.0698	0.0913
28	0.0384	1.1510	0.0094	0.0038	0.0035	-0.0026
29	0.0905	1.1642	-0.2079	0.0652	0.0547	-0.1631
30	0.0511	1.1508	0.0845	0.0635	0.0121	-0.0531
31	0.0557	1.1661	-0.0557	-0.0350	0.0373	0.0031
32	0.0678	1.1461	-0.1586	-0.0472	-0.1025	0.0636
33	0.0996	1.2116	0.1252	0.0015	0.1025	-0.0256

Sum of Residuals	0
Sum of Squared Residuals	5.0533
Predicted Resid SS (Press)	6.2242

[mlrplus-1] The SAS System

Correlation Analysis

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 31

	X1	X2	X3	X4	X5
X1	1.00000 0.0	0.33287 0.0673	0.22262 0.2287	0.09740 0.6022	-0.84525 0.0001
X2	0.33287 0.0673	1.00000 0.0	-0.18150 0.3285	0.17871 0.3361	-0.25426 0.1675
X3	0.22262 0.2287	-0.18150 0.3285	1.00000 0.0	0.21952 0.2354	-0.03154 0.8662
X4	0.09740 0.6022	0.17871 0.3361	0.21952 0.2354	1.00000 0.0	0.34868 0.0546
X5	-0.84525 0.0001	-0.25426 0.1675	-0.03154 0.8662	0.34868 0.0546	1.00000 0.0

[mlrplus-2] The SAS System

Dependent Variable: X5 < 독립변수가 4개인 경우 >

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	4	12.85592	3.21398	63.943	0.0001
Error	26	1.30685	0.05026		
C Total	30	14.16277			

Root MSE 0.22419 R-square 0.9077

Parameter Estimates

Variable	DF	Parameter Estimate	T for H0: Parameter=0	Prob > |T|	Standardized Estimate
INTERCEP	1	1.954229	2.543	0.0173	0.00000000
X1	1	-0.162704	-13.498	0.0001	-0.89624893
X2	1	-0.008802	-0.276	0.7848	-0.01862465
X3	1	0.123689	1.087	0.2871	0.07160966
X4	1	0.240712	6.748	0.0001	0.42358097

Dependent Variable: X5 < 독립변수가 2개인 경우 >

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	12.77474	6.38737	128.849	0.0001
Error	28	1.38803	0.04957		
C Total	30	14.16277			

Root MSE 0.22265 R-square 0.9020

Parameter Estimates

Variable	DF	Parameter Estimate	T for H0: Parameter=0	Prob > |T|	Standardized Estimate
INTERCEP	1	2.579694	25.031	0.0001	0.00000000
X1	1	-0.161140	-14.932	0.0001	-0.88763217
X4	1	0.247277	7.320	0.0001	0.43513321

[mlrplus-3] The SAS System

Obs	Dep Var X5	Predict Value	Std Err Predict	Lower95% Mean	Upper95% Mean	Lower95% Predict	Upper95% Predict
1	2.1200	2.0085	0.055	1.8962	2.1208	1.5388	2.4782
2	3.3900	3.1604	0.058	3.0414	3.2793	2.6891	3.6317
3	3.6100	3.7427	0.088	3.5625	3.9230	3.2523	4.2331
4	1.7200	1.6479	0.083	1.4775	1.8183	1.1610	2.1347
5	3.2100	3.3454	0.081	3.1798	3.5111	2.8602	3.8306
6	2.5900	2.3180	0.042	2.2315	2.4045	1.8538	2.7822
7	3.2500	2.9626	0.056	2.8486	3.0765	2.4925	3.4327
8	2.8600	2.8884	0.058	2.7695	3.0073	2.4171	3.3597
9	2.3200	2.1457	0.089	1.9630	2.3285	1.6544	2.6371
10	1.5700	1.8457	0.094	1.6534	2.0380	1.3507	2.3407
11	1.5000	1.7229	0.063	1.5930	1.8528	1.2487	2.1971
12	2.6900	2.6147	0.049	2.5143	2.7152	2.1477	3.0818
13	4.0600	3.8049	0.093	3.6146	3.9953	3.3107	4.2992
14	1.9800	2.4680	0.094	2.2759	2.6601	1.9731	2.9629
15	2.2900	2.2550	0.069	2.1140	2.3959	1.7776	2.7323
16	3.5500	3.4460	0.081	3.2805	3.6114	2.9608	3.9311
17	3.3100	3.1843	0.077	3.0265	3.3421	2.7017	3.6669
18	1.8300	1.6623	0.092	1.4740	1.8506	1.1689	2.1557
19	1.6900	1.5618	0.072	1.4142	1.7094	1.0824	2.0411
20	2.4200	2.5286	0.041	2.4453	2.6119	2.0650	2.9922
21	2.9800	2.8620	0.074	2.7110	3.0130	2.3816	3.3424
22	1.8400	2.1696	0.052	2.0631	2.2762	1.7013	2.6380
23	2.4800	2.5661	0.048	2.4681	2.6641	2.0996	3.0326
24	2.4100	2.3044	0.073	2.1547	2.4542	1.8244	2.7844
25	1.7800	2.0827	0.049	1.9819	2.1834	1.6156	2.5498
26	2.2200	2.1569	0.045	2.0644	2.2494	1.6915	2.6222
27	2.7200	3.1101	0.069	2.9690	3.2512	2.6327	3.5875
28	2.3600	2.1696	0.052	2.0631	2.2762	1.7013	2.6380
29	2.8100	3.0120	0.055	2.8993	3.1247	2.5422	3.4818
30	1.6400	1.7604	0.061	1.6363	1.8845	1.2877	2.2331
31	1.8200	1.5123	0.073	1.3620	1.6627	1.0321	1.9925

[mlrplus-4] The SAS System

Obs	Residual	Std Err Residual	Student Residual	-2-1-0 1 2	Cook's D	Rstudent
1	0.1115	0.216	0.517	*	0.006	0.5098
2	0.2296	0.215	1.068	**	0.028	1.0711
3	-0.1327	0.205	-0.649	*	0.026	-0.6421
4	0.0721	0.207	0.349		0.007	0.3436
5	-0.1354	0.207	-0.653	*	0.022	-0.6460
6	0.2720	0.219	1.244	**	0.019	1.2571
7	0.2874	0.216	1.333	**	0.039	1.3529
8	-0.0284	0.215	-0.132		0.000	-0.1297
9	0.1743	0.204	0.854	*	0.047	0.8500
10	-0.2757	0.202	-1.366	**	0.134	-1.3882
11	-0.2229	0.213	-1.044	**	0.032	-1.0461
12	0.0753	0.217	0.347		0.002	0.3410
13	0.2551	0.202	1.261	**	0.112	1.2746
14	-0.4880	0.202	-2.417	****	0.420	-2.6676
15	0.0350	0.212	0.166		0.001	0.1626
16	0.1040	0.207	0.501	*	0.013	0.4945
17	0.1257	0.209	0.602	*	0.016	0.5948
18	0.1677	0.203	0.827	*	0.047	0.8221
19	0.1282	0.211	0.609	*	0.014	0.6018
20	-0.1086	0.219	-0.496		0.003	-0.4893
21	0.1180	0.210	0.562	*	0.013	0.5546
22	-0.3296	0.216	-1.523	***	0.045	-1.5613
23	-0.0861	0.217	-0.396		0.003	-0.3899
24	0.1056	0.210	0.502	*	0.010	0.4953
25	-0.3027	0.217	-1.394	**	0.033	-1.4189
26	0.0631	0.218	0.290		0.001	0.2848
27	-0.3901	0.212	-1.843	***	0.120	-1.9302
28	0.1904	0.216	0.879	*	0.015	0.8757
29	-0.2020	0.216	-0.936	*	0.019	-0.9343
30	-0.1204	0.214	-0.562	*	0.008	-0.5550
31	0.3077	0.210	1.464	**	0.087	1.4958

[mlrplus-5] The SAS System

Obs	Hat Diag H	Cov Ratio	Dffits	INTERCEP Dfbetas	X1 Dfbetas	X4 Dfbetas
1	0.0606	1.1536	0.1295	0.0824	0.0453	-0.0801
2	0.0680	1.0562	0.2894	0.0923	-0.1963	0.0928
3	0.1562	1.2630	-0.2762	-0.0643	0.2294	-0.1108
4	0.1395	1.2793	0.1383	-0.0572	0.1118	0.0360
5	0.1319	1.2269	-0.2518	0.0721	0.1034	-0.2020
6	0.0360	0.9754	0.2430	0.1326	0.0368	-0.0725
7	0.0625	0.9771	0.3492	0.2507	-0.2307	-0.0527
8	0.0680	1.1944	-0.0350	-0.0289	0.0216	0.0112
9	0.1605	1.2274	0.3717	0.3544	-0.0358	-0.3253
10	0.1778	1.1031	-0.6456	0.3838	-0.4440	-0.3345
11	0.0811	1.0773	-0.3107	-0.0040	-0.2397	0.0493
12	0.0485	1.1572	0.0770	-0.0142	0.0038	0.0439
13	0.1742	1.1334-	0.5855	0.2231	-0.5201	0.1438
14	0.1774	0.6721	-1.2389	-1.2350	0.3988	1.0035
15	0.0955	1.2295	0.0528	-0.0270	0.0255	0.0320
16	0.1316	1.2500	0.1925	0.1254	-0.1648	-0.0125
17	0.1198	1.2185	0.2194	-0.0780	-0.0638	0.1817
18	0.1704	1.2483	0.3727	0.2934	0.0964	-0.3293
19	0.1047	1.1969	0.2058	-0.0097	0.1705	-0.0317
20	0.0333	1.1236	-0.0909	-0.0490	0.0113	0.0107
21	0.1096	1.2108	0.1946	-0.0946	-0.0022	0.1629
22	0.0546	0.9102	-0.3752	-0.2821	-0.0604	0.2371
23	0.0462	1.1498	-0.0858	-0.0709	0.0255	0.0370
24	0.1078	1.2166	0.1722	-0.0953	0.0766	0.1140
25	0.0488	0.9450	-0.3214	-0.1756	-0.1188	0.1554
26	0.0411	1.1527	0.0590	0.0246	0.0225	-0.0178
27	0.0957	0.8369	-0.6280	0.1765	0.1954	-0.4894
28	0.0546	1.0846	0.2104	0.1582	0.0339	-0.1330
29	0.0611	1.0797	-0.2383	-0.1498	0.1621	0.0071
30	0.0740	1.1641	-0.1569	-0.0477	-0.1030	0.0671
31	0.1087	0.9854	0.5223	0.0139	0.4295	-0.1271

Sum of Residuals	0
Sum of Squared Residuals	1.3880
Predicted Resid SS (Press)	1.7596

4.5 변수선택

모형선택(model selection), 또는 **변수선택**(variable selection)의 문제는, 수집된 모든 회귀변수들중 일부, 즉, 부분집합만으로 반응변수에 대한 설명을 하려고 시도하는데서 제기된다. 일반적으로 어떤 조사에서건 반응변수의 원인이 될만한 회귀변수의 수는 적지않을 것이며, 이들 가운데에서 반응변수의 변화 유형을 가장 잘 설명할 수 있는 회귀변수들을 추리는 것이 경제적인 통계분석일 것이다.

회귀변수가 p 개 있다면, 이 회귀변수들의 부분집합으로 만들 수 있는 회귀모형의 가짓수는 2^p 개가 된다. 만일 $p=10$이면 가능한 회귀모형의 갯수는 약 1000 개가 되며, 만일 $p=20$이면 가능한 회귀모형의 개수가 약 1000000 개에 이르게 된다. 이들중 도대체 어느 모형이 '가장 좋은 모형' (best model)인지 찾는 문제는 그리 간단하지 않다. 왜냐하면, 두 모형을 서로 비교하려면 일단 회귀선을 적합한 후, 회귀계수들이 제대로 해석되는지, 반응변수에 대한 추측은 제대로 되는지, 과연 적합된 회귀선이 통계적으로 유의한지 등을 비교 검토해야 하기 때문이다. 게다가, 방금 언급한 문제점들은 따로따로 독립적이지 않고 서로 맞물려있으며, 회귀변수 간에 공선성이 존재하면 모형선택의 문제는 더욱 어려워진다.

게다가, 두 회귀선 가운데 하나를 선택해야한다면, 도대체 어떤 통계적 판정기준(criterion)을 사용하여 비교해야할까? 앞 단락에서 좋은 모형을 강조한 것은 바로 이런 이유에서이다. 일례로 결정계수를 판정기준으로 삼아 좋은 모형을 찾는다면 언제나 완전 모형(full model)이 선택될 것이다. 왜냐하면, 결정계수는 회귀변수가 추가될 때마다 그 값이 항상 커지기 때문이다. 그간 제안된 좋은 모형을 선택하는 판정기준은 부지기수로 많으며, 각 판정기준으로 선택되는 가장 좋은 모형이 전부 다 다를 수도 있다.

모형선택법의 기초적인 아이디어는 1960년대에 주로 연구되었는데 거기에는 그럴만한 이유가 있다. 예를 들어, 회귀변수의 갯수가 10일 때, 모든 가능한 회귀모형을 적합한 후 비교하려면, 앞서도 언급하였지만, 모두 약 1000 번의 회귀분석을 해야한다는 결론에 이르게 되며, 이 별로 크지않은 숫자가 범용 컴퓨터의 초창기인 1960 년대의 통계학자들에게는 얼마나 터무니없는 숫자였는지는 짐작이 갈 것이다. 따라서, 그 당시의 학자들은 모든 가능한 회귀 모형들을 적합하지 않으면서도 최적의 모형을 찾을 수 있는 방법들을 개발하였으며, 여기에 속하는 모형선택법으로 전방선택, 후방소거, 단계별회귀 등이 있다. 이 방법들은 기본적으로 각 모형에서 계산되는 결정계수를 판정기준 통계량으로 사용하여 최적의 회귀변수들을 찾는다. 그러나, 최근에는 컴퓨터 계산 능력의 급속한 발전으로 말미암아 회귀 변수의 갯수가 최대 13 개 정도까지는 모든 가능한 회귀모형을 적합시키고, 동시에 여러가지 판정기준 통계량을 출력하여도 그다지 시간이 많이 걸리지 않게 되었으므로, 학자에 따라서는 앞에 언급한 여러 방법들은 모두 구식으로 치부하기도 한다.

하지만, 사회과학 분야의 조사결과 등을 보면 회귀변수의 갯수를 수십개 정도 잡는 일은 비일비재하고, 이 정도의 갯수라면 모든 가능한 회귀모형의 적합이란, 아무리 현재의 컴퓨터가 발전되었고 효율적인 계산 앨고리듬을 갖고있어도, 그다지 쉬운 일은 아니다. 덧붙여, 결과되는 엄청난 컴퓨터 출력을 체계적으로 검사하여 최적이라 판단되는 모형을 찾는 작업은 거

의 불가능에 가깝다. 이런 상황이라면 앞 단락의 오래된 방법들을 쓸 수 밖에 없을 것이다.

(1) R^2

이 방법은 모든 회귀변수들을 기초로, 모든 가능한 회귀모형을 전부 적합한 후, 각 모형에 대한 R^2 값을 출력한다. 이 방법은 사용하면 주어진 갯수의 회귀변수들에 대한 모형 중 어느 것이 가장 R^2 값을 크게하는지는 알 수 있으나, 그 모형이 가장 최상의 모형이란 보증이 없음을 유의하라.

(2) 맬로우스의 C_P

맬로우스(Mallows)의 C_p 판정기준은 적합된 모형의 추측 능력을 수량화하여 최적의 모형을 선택하는 방법으로 많이 사용되고 있다. 이 판정 기준에서는 **C_p 값이 p(회귀변수의 갯수)와 거의 같은 모형을 최고라고 판단하며, 가급적 p가 작은 모형을 찾아야한다.** 예시한 출력에서, In은 모형에 포함된 회귀변수의 갯수이며, 이 값에 대응되는 C(p) 값과 거의 같아야 좋은 모형이다. 출력을 보면, 이 두 값의 차이가 적은 순서대로 나열되어있다. 따라서, C_p 기준 하에서는 회귀변수 X_4와 X_5만을 포함하는 회귀모형이 가장 좋다고 할 수 있다. C_p 값의 출력은 전방선택, 후방소거, 단계별회귀 등에서도 분산분석표 상단에 출력되는데, 그와 같은 절차들에서 최적으로 판정하였어도 C_p 값이 최적이 아니면 좋은 모형이라 할 수 없다. C_p 판정기준을 선택하면 자동적으로 결정계수도 출력된다.

SAS

N = 22 Regression Models for Dependent Variable: Y

	C(p)	R-square	In	Variables in Model
==>	2.14558	0.95841490	2	X4 X5
	2.55650	0.95747318	2	X1 X5
	2.73542	0.96164663	3	X2 X4 X5
	3.81376	0.95917535	3	X1 X4 X5
	4.09509	0.95853062	3	X3 X4 X5
	4.15247	0.95839911	3	X1 X2 X5
	4.15818	0.96296953	4	X1 X2 X4 X5
	4.55633	0.95747357	3	X1 X3 X5
	4.60369	0.96194853	4	X2 X3 X4 X5
	5.75562	0.95930858	4	X1 X3 X4 X5
	5.77941	0.94550358	1	X5
	5.98585	0.95878095	4	X1 X2 X3 X5
	6.00000	0.96333203	5	X1 X2 X3 X4 X5
	6.41721	0.94862539	2	X2 X5
	6.83166	0.94767559	2	X3 X5
	8.32649	0.94883330	3	X2 X3 X5
	10.78459	0.93861647	2	X1 X3
	11.32754	0.94195565	3	X1 X2 X3
	12.78209	0.93862220	3	X1 X3 X4
	12.79506	0.94317596	4	X1 X2 X3 X4
	15.08193	0.92876803	2	X3 X4
	15.90841	0.92229045	1	X1
	15.93250	0.93140225	3	X2 X3 X4
	17.65245	0.92287707	2	X1 X2
	17.76193	0.92262615	2	X1 X4
	19.53347	0.91398274	1	X3
	19.62537	0.92293911	3	X1 X2 X4
	20.35534	0.91668271	2	X2 X3
	29.38906	0.89597971	2	X2 X4
	31.86208	0.88572867	1	X4
	33.82441	0.88123150	1	X2

4.6 추가 예제

데이타: 지능지수, 키, 무게
목표: 지능지수를 키와 무게로 설명해보자.

IQ	무게(*kg*)	키(*cm*)
150	76	178
140	70	170
127	50	160
145	70	180
100	90	165
110	88	165
80	74	181

① 단위: *kg*과 *cm*

IQ = 109.7 − 0.73WT(*kg*) + 0.38HT(*cm*) (p 값 = 0.731, R^2 = 0.145)
표준화 계수: WT = −0.37, HT = +0.12

② 단위: *g*과 *m*

IQ = 109.7 − 0.00073WT(*g*) + 38HT(*m*) (p 값 = 0.731, R^2 = 0.145)
표준화 계수: WT = −0.37, HT = +0.12

③ 반복자료: 동일 자료를 10회 반복 입력함

IQ = 109.7 − 0.73WT(*kg*) + 0.38HT(*cm*) (p 값 = 0.005^{**}, R^2 = 0.145)
표준화 계수: WT = −0.37, HT = +0.12

④ outlier를 제거하면

IQ = −157.6 − 0.83WT(*kg*) + 2.05HT(*cm*) (p 값 = 0.03^{*}, R^2 = 0.9)
표준화 계수: WT = −0.6, HT = +0.8

※ 교훈

▶ 변수의 상대적 중요도는 표준화 회귀계수로 사정한다. (공선성이 없을 때에)
▶ 모형의 유의성과 적합도(설명력)는 항상 일치하지 않는다.

▶ p 값과 R^2 값을 기초로 적합도를 판정해야 한다.
▶ outlier(이상점)가 반드시 나쁜 점이라고 할 수 없다.
▶ 모형의 유의성에 크게 집착할 필요가 없다.
▶ 유의수준 α의 값을 항상 0.05로 놓아야 되는 것은 아니다.
▶ 상관관계 ≠ 인과관계

4.7 Cronbach's alpha

크론박의 알파 계수는 동일 특성을 측정하는 여러 문항의 내적 합치도, 또는 문항들의 동등성에 대한 신뢰도를 재는 데 사용한다. 예를 들어 어떤 설문지에서 문항 1, 2, 3이 직업에 대한 만족도를 묻고있다고 가정하자. 또 문항마다 5점 척도를 사용하고 척도 순서의 방향이 일치한다고 하자. 문항들의 반응값을 변수 q1, q2, q3로 받는다면 이때 크론박의 알파는 다음과 같이 CORR 절차에서 출력된다. SAS 출력의 기본 형식은 다음과 같다.

The SAS System

CORRELATION ANALYSIS

Cronbach Coefficient Alpha

for RAW variables : 0.6552
for STANDARDIZED variables: 0.7088

	RAW Variables		Std. Variables	
Deleted Variable	Correlation With Total	Alpha	Correlation with Total	Alpha
Q1	0.5461	0.8206	0.4986	0.8933
Q2	0.7331	0.6480	0.6102	0.6860
Q3	0.9107	0.4086	0.9020	0.4621

출력을 보면 문항 1을 뺄 경우 알파값이 0.6559에서 0.8206으로 증가하므로 신뢰도를 크게 하려면 문항 1을 제외하는 것이 바람직하다. 문항 3을 빼면 알파값이 0.4086으로 감소하므로 문항 3은 제외될 수 없다.

알파 계수는 회귀분석에서 결정계수와 비슷한 역할을 한다. 즉, 회귀분석에서 설명변수의 갯수가 늘어나면 늘어날수록 결정계수가 증가하듯, 문항수를 늘리면 알파값도 증가한다. 그러나 무작정 문항수를 늘리는 것은 바람직하지 않다. 알파값이 큰 것이 물론 좋지만 작다고 해서 분석이 불가능한 것은 아니다.

5. 비교실험 분석법에 관하여

▶ 관찰연구(observational study), 또는 조사연구(survey)

목적: 주어진 현상의 관찰
예: 사회과학분야의 설문조사
자료: 표본조사론(sampling theory)을 기초로 수집
분석: 그래프, 줄기-잎 그림, 산점도, 상관분석, 회귀분석, 교차표 등

▶ 비교실험(comparative experiments), 또는 간섭연구(interference study)

목적: 인과(cause and effect) 관계의 탐지
예: 물리 및 생물과학의 실험
자료: 실험계획법(design of experiments)을 기초로 수집
분석: 관찰연구의 분석법을 포함하여,
분산분석(analysis of variance), 반응표면분석(response surface analysis),
공분산분석(analysis of covariance), 다변량분석(multivariate analysis),
비모수 통계(nonparametric statistics) 등

▶ 어떻게 하면 실험을 잘 할 수 있을까?

최소의 노력(비용, 시간, 실험횟수)으로 최대 효과(정보)를 얻자. → 실험계획법

▶ 용어

요인(factor): 3개 이내로.
수준(level): 요인에 대한 조절 조건
처리(treatment): 요인 수준들의 조합
반응(response): 관측값(들) (동시에 측정되는 반응이 둘 이상이면 다변량자료)

▶ 비교실험의 통계적 해석을 가능케하는 근본 원리

① 랜덤화(randomization)
실험의 객관성을 보장. 실험 순서를 랜덤하게 결정하는 것.

② 반복(replication)
동일 처리에서 두 번 이상의 실험을 함. 반복 실험을 하면 오차의 수량화가 가능

③ 블록화(blocking)

실험의 정밀도를 증가시킨다. 필요하다고 판단되면 블록화는 하는 편이 유리하다.

▶ 실험자는 모름지기 자신이 수행하는 실험에 대한 명확한 아이디어가 있어야 하며 어떠한 제한점이 있는지 등을 누구에게나 말할 수 있어야 한다. 요인, 요인 수준, 반응값의 선택에 있어서는 경험적이며 비통계적인 지식이 필수적이다. 결과 해석 역시 통계분석에만 의존할 필요가 없다. 복잡하고 정교한 실험과 고차원적인 분석법을 사용하는 것이 반드시 바람직하지는 않다. 실험계획과 분석은 간단하면 간단할수록 최상이라 할 수 있다. 실험이 복잡하면 결과 해석이 더 어려워지고 적합한 통계분석법이 존재하지 않는 경우가 많게 된다. 통계학이 절대 만병통치약은 아니다. 통계적인 유의성과 실용적인 유의성이 반드시 합치되지 않음도 인식해야 한다. 예를 들어, 빵을 굽는 최적의 조건을 찾아냈다 하더라도 제빵기를 교체하는데 드는 비용이 막대하면 교체하지 않는 선에서 최적 조건을 찾아볼 필요가 있는 것이다. 또한 실험은 한 번으로 끝나지 않는다는 점을 명심해야 한다. 단 한 번의 실험으로 반응값에 가장 영향을 주는 요인과 요인의 수준이 찾아지지는 않는다. 일차실험을 한 후 결과를 보아 이차실험, 삼차실험 등을 계획하노라면 바라는 결과를 얻게 되기 마련이다.

▶ 선형 모형을 사용하는 모수(parametric) 통계 분석법의 기초 가정

단일 표본에서 오차들은 서로 독립이고 분산이 동일한 정규분포를 따른다.

과연 그럴까? **그렇다고 하는 것이다!**

데이타를 대표하는 가장 적절한 통계량은 평균이다.

과연 그럴까? **이것을 부인하면 가용한 분석법이 별로 없다.**

※ 우리가 설정한 모형이 과연 합당한가?

실험을 어떻게 계획하느냐가 모형을 결정한다. 회귀분석에서는 더 나은 모형으로 바꿀 수 있지만 분산분석에서는 모형의 변경이 불가능하다.

▶ 비모수 통계 분석법의 기초 가정

자료의 분포로 정규분포나 이항분포와 같은 특정 분포를 가정하지 않고 단순히 연속분포나 이산분포를 가정한다. 자료의 분포에 대하여 최소한의 가정을 하는 이점이 있으나 가용한 분석법이 한정되어 있고, 모수적 추론이 가능할 때 비모수 기법을 사용하면 정보의 손실을 초래한다. 비모수 통계에서는 대표값으로 흔히 중앙값(median)을 사용한다.

▶ 비교 분석법의 유형

통계적 자료분석에서는 처리간 대표값(위치모수; location parameter)을 비교하여 효과 유무를 판정한다. 모수적 추론일 때는 위치모수로 평균, 비모수적 추론에서는 위치모수로 중앙값을 사용하고, 원시 자료대신 주로 순위(rank) 자료를 사용한다. 아래 표의 가설에서 모평균은 μ, 모중앙값은 m으로 표기하였다.

상황		모수적 추론		비모수적 추론	
		귀무가설	분석법	귀무가설	분석법
단일 표본		$\mu=\mu_0$	t 검증	$m=m_0$	부호검증, 윌콕슨 부호순위검증
이표본	독립표본	$\mu_1=\mu_2$	합동 t 검증	$m_1=m_2$	맨-휘트니 U 검증, 윌콕슨 순위합 검증
	짝지어진 표본	$\mu_1-\mu_2=\mu_d=0$	t 검증	$m_1-m_2=m_d=0$	부호검증, 윌콕슨 부호순위검증
삼표본 이상		$\mu_1=\mu_2=\cdots=\mu_k$	분산분석, F 검증	$m_1=m_2=\cdots=m_k$	크루스칼-월리스 검증

① 단일표본(single sample): 자료의 대표값과 가정된 값 간의 비교

분석: t 검증, 부호검증(sign test), 윌콕슨 부호 순위 검증(Wilcoxon signed-rank test), SAS/UNIVARIATE
자료: 어떤 품종의 말의 수명
질문: 말의 평균 수명을 22년이라 할 수 있는가?

말의 수명				
17.2	18.0	18.7	19.8	20.3
20.9	21.0	21.7	22.3	22.6
23.1	23.4	23.8	24.2	24.6
25.8	26.0	26.3	27.2	27.6
28.1	28.6	29.3	30.1	35.1

② 두 표본(two sample): 두 자료의 대표값 간의 비교

(i) 독립표본(independent samples)

분석: 이표본 t 검증, SAS/TTEST; Mann-Whitney U 검증, 또는 Wilcoxon 순위합(rank sum) 검증, SAS/NPAR1WAY
자료: 두 종류의 지혈제의 효과를 측정. 실험 참여자의 살갗에 작은 상처를 내어 피

가 흐르게 한 다음 지혈제를 투여하여 완전히 지혈이 될 때까지의 시간(단위: 분)을 측정하였다.

질문: 지혈제 G가 지혈제 B보다 더 효과적인가?

지혈제 B	지혈제 G
8.8 8.4 7.9 8.7 9.1 9.6	8.8 9.9 9.0 11.1 9.6 8.7 9.5 10.4

(ii) 짝지어진 표본 – 가장 간단한 블록화

분석: t 검증, 부호검증(sign test), 윌콕슨 부호 순위 검증(Wilcoxon signed-rank test), SAS/UNIVARIATE

자료: 사슴의 왼쪽 뒷다리와 왼쪽 앞다리의 길이

질문: 다리 길이가 같은가?

사슴	뒷다리	앞다리
1	142	138
2	140	136
3	144	147
4	144	139
5	142	143
6	146	141
7	149	143
8	150	145
9	142	136
10	148	146

한 사슴에 대하여 두 다리 길이를 측정하였으므로 사슴들은 독립이지만 한 마리 사슴의 일부인 두 다리는 독립이 아니다.

③ 분산분석(anova): 셋 이상의 그룹 대표값(평균) 간의 비교

(i) 일원분산분석(one-way anova)

구분: 요인이 하나. 수준이 셋 이상.

분석: F 검증, SAS/GLM, ANOVA; Kruskal-Wallis 검증, SAS/NPAR1WAY

자료: 세 품종의 사과나무 잎에 함유된 인의 함량. 각 품종마다 5 장의 잎을 랜덤 추출하여 인의 함량을 측정.

질문: 품종 간 인의 함량에 차이가 나는가?

품종		
1	2	3
0.35	0.65	0.60
0.40	0.70	0.80
0.58	0.90	0.75
0.50	0.84	0.73
0.47	0.79	0.66

(ii) 이원분산분석(two-way anova)

구분: 요인이 둘. 수준 조합마다 실험.
분석: F 검증, SAS/GLM, ANOVA
자료: 4 종류의 살충제가 3 가지 품종의 밀감 나무의 생산량에 미치는 영향을 알기 위하여, 밀감 농장에서 각 품종별로 8 그루씩의 나무를 랜덤으로 선택하고, 각 살충제마다 2 그루씩 랜덤으로 배정한다. 규정된 사용법에 따라 살충제를 살포하고, 일정 기간 후 나무에 달린 밀감의 숫자를 헤아렸다.
질문: 살충제 간에 차이가 나는가? 품종 간 차이는 있는가? 살충제와 품종 간에 상호작용이 존재하는가?

품종	살충제							
	A		B		C		D	
1	49	39	50	55	43	38	53	48
2	55	41	67	58	53	42	85	73
3	66	68	85	92	69	62	85	99

(ii) 이원분산분석: 블록 계획

구분: 관심 요인은 하나. 블록 요인을 삽입.
분석: F 검증, SAS/GLM, ANOVA; Friedman 검증
자료: Guinea pig를 사용하여 네 가지 다이어트에 대한 체중감소 효과를 측정. 블록 요인은 쥐가 갇혀있는 우리로, 각 우리마다 환경 조건, 이를테면, 채광도, 소음, 온도, 습도 등을 서로 다르게 설정하였다.
질문: 다이어트 간에 차이가 나는가?

우리(블록)	다이어트			
	1	2	3	4
1	1.5	2.7	2.1	1.3
2	1.4	2.9	2.2	1.0
3	1.4	2.1	2.4	1.1
4	1.2	3.0	2.0	1.3
5	1.4	3.3	2.5	1.5

(iii) 분할구계획(split-plot design): 블록화, repeated measures

구분: 이중 블록화. 셋 이상의 시간대에서 측정해도 분석 가능.
분석: F 검증, SAS/GLM (SAS에서는 다변량 분석으로 처리함)
자료: 대학생들의 신체 강도. 학생들을 세 그룹으로 나눠 신체 강도를 미리 측정한 다음, 그룹 1과 2에는 새로 고안된 두 가지의 신체 단련 체조를 가르쳤고, 그룹 3은 조절 그룹으로 아무런 체조도 시키지않았다. 그 후, 한 달 뒤와 두달 뒤에 다시 신체 강도를 측정한 결과이다.
질문: 그룹 간에 차이가 나는가? 교습 전후에 차이가 나는가? 상호작용은 있는가?

그룹 1				그룹 2				그룹 3			
학생	선측	후측1	후측2	학생	선측	후측1	후측2	학생	선측	후측1	후측2
1	26.25	29.50	29.50	8	27.47	28.74	28.74	15	22.27	22.52	22.52
2	24.33	27.62	27.62	9	25.19	26.11	26.11	16	21.55	21.79	21.79
3	22.52	25.71	25.71	10	23.53	25.45	25.45	17	23.31	23.53	23.53
4	29.33	31.55	31.55	11	24.57	25.58	25.58	18	30.03	30.21	30.21
5	28.90	31.35	31.35	12	26.88	27.70	27.70	19	28.17	28.65	28.65
6	25.13	29.07	29.07	13	27.86	28.82	28.82	20	28.09	28.33	28.33
7	29.33	31.15	31.15	14	28.09	28.99	28.99	21	27.55	27.86	27.86

④ 공분산분석(anocova)

구분: 반응값의 추이를 공변수로 설명한 다음 처리 간 효과 차이를 비교하자.
분석: F 검증, SAS/GLM
자료: 아홉 명의 학생을 랜덤으로 세 그룹으로 나누었다. 각 그룹마다 동일한 시험을 치고 선시험(pre-test) 성적을 기록한다. 그 다음, 세 가지 서로 다른 교육 방법으로 동일한 내용을 가르친 후, 다시 시험을 치루고 후시험(post-test) 성적을 기록한다.
질문: 교육 방법에 차이가 있는가?
접근: 개인 차이가 문제될 수 있다. 후시험 성적에 끼치는 선시험 성적의 영향을 회

귀분석으로 제거한 후 분산분석을 하자. 공분산분석은 가정이 많은 기법이므로 사용상 점검할 부분이 많다. 선시험 성적이 공변수이며, 공변수의 갯수는 둘 이상이어도 상관없다. 즉, 후시험 성적에 영향을 끼칠 수 있는 다른 공변수로 지능지수를 고려할 수도 있다. 그러나 셋 이상의 시간대에서 반복 측정한 자료라면 공분산분석을 적용 못한다.

그룹 1		그룹 2		그룹 3	
선시험	후시험	선시험	후시험	선시험	후시험
29	39	17	35	1	38
4	34	35	38	15	43
18	36	3	32	32	44

▶ 기타 분석

처리 수준에 따른 반응값들의 변화 패턴을 회귀 곡선으로 적합할 수 있다.
두 요인 간의 변화 패턴의 동질성 유무를 카이제곱 검증법으로 탐색할 수 있다.

▶ 기타 사항

고정효과냐 랜덤효과냐에 따라 요인 효과의 분석과 해석이 달라진다.

▶ 실험의 진행 방법에 따라 모형과 분석이 달라진다.

6가지의 인절미를 관능 평가하는 상황을 생각해보자.

6. 다변량해석 개요

동일 개체에 대하여 측정한 반응값이 동시에 둘 이상인 자료를 다변량 자료라 부른다. 다음 자료를 보자. 20 명의 여성에 대하여 Sweat rate, Sodium content, Potassium content의 세 값을 측정하였다.

여성	Sweat	Sodium	Potassium
1	3.7	48.5	9.3
2	5.7	65.1	8.0
3	3.8	47.2	10.9
⋮	⋮	⋮	⋮
20	5.5	40.9	9.4

이때 (Sweat, Sodium, Potassium)=(4, 50, 10)이라고 놓을 수 있는지 유무를 추론한다면 이것이 곧 다변량 자료에서 평균벡터에 대한 추론이 된다. 만일 이 자료에 대하여 단일변량해석을 한다면 Sweat=4, Sodium=50, Potassium=10이라는 세 개의 독립적인 가설을 따로따로 검증해야 한다. 다변량 해석에서는 단일변량해석과 달리 동시에 세 변수에 대한 추론을 전개함을 유념하라.

분산분석 스타일의 실험에서 동시에 측정되는 반응변수가 여러 개일 때의 분석을 다변량 분산분석(manova: multivariate anova)이라 부른다. 이 분석은 SAS/GLM으로 결과를 출력시킬 수 있다.

그러나, 사회경제분야에서는 다변량 분산분석보다는 주성분 분석(principal components analysis), 인자 분석(factor analysis), 정준상관분석(canonical correlation analysis)과 같은 공분산 구조의 분석과 판별분석(discriminant analysis), 군집분석(cluster analysis)과 같은 분류(classification) 기법이 더 많이 사용된다. 기타 다변량 회귀분석을 기초로 하는 경로분석(path analysis) 등이 때로 사용된다.

(1) 주성분분석

주성분분석에서는 자료의 공분산 구조를 두세개의 원래 변수들의 선형조합을 기초로 설명하는 것을 목표한다. 이렇게 되면 자료의 축소와 해석의 편의를 얻을 수 있다. 주성분분석은 자료의 공분산행렬을 기초로 수행하기도 하고 상관행렬을 기초로 수행하기도 한다. 반응값들의 척도가 다르거나, 같더라도 변동이 크다면 상관행렬을 기초로 주성분분석을 해야한다.

다음 데이타를 보자.

< Stock-price data: weekly rate of return >

Week	Allied Chemical z_1	du Pont z_2	Union Carbide z_3	Exxon z_4	Texaco z_5
1	.000000	.000000	.000000	.039473	.000000
2	.027027	-.044855	-.003030	-.014466	.043478
⋮	⋮	⋮	⋮	⋮	⋮
100	.019108	-.033303	.008362	.033898	.004566

이 주식시세 자료에 주성분분석을 적용하는 예를 들자. 5×5의 상관행렬을 기초로 분석을 진행하면 다음 두 개의 선형조합이 자료변동의 약 73%를 설명하고 있음을 알 수 있다.

$$\widehat{y_1} = .464z_1 + .457z_2 + .470z_3 + .421z_4 + .421z_5 \quad (57.14\%)$$

$$\widehat{y_2} = .240z_1 + .509z_2 + .260z_3 - .526z_4 - .582z_5 \quad (16.18\%)$$

첫번째 성분의 계수들은 동일 부호의 비슷한 값을 갖는다. 따라서 이 성분은 시장 상황을 반영하는 시장 성분(market component)이라 할 수 있다. 두번째 성분을 보면 화학회사와 정유회사 간 부호 차이가 두드러진다. 따라서 이는 업종의 차이라 할 수 있고, 산업 성분(industry component)이라 부를 수 있겠다. 결국 주식시세는 전반적인 시황과 각 산업에 고유한 활동에 영향을 받는다고 하겠다.

(2) 인자분석

인자분석은 주성분분석의 확장판으로 회전시키지 않은 직교인자모형은 본질적으로 주성분분석과 같다. 그러나, 인자분석에서는 관측될 수 없는 변량(인자)들의 선형조합으로 공분산구조를 설명한다. 원리는 다음과 같다. 많은 변수들이 상관도로 그룹화된다고 가정하자. 즉, 어느 그룹 내의 변수들은 서로 상관도가 높으나 다른 그룹들의 변수들과는 상대적으로 낮은 상관도를 갖고 있다. 이 경우 각 그룹을 전체 상관행렬을 설명하는 인자(factor)로 파악할 수 있다. 인자분석에서는 가능하면 적은 수의 인자로 공분산구조를 설명하려 한다. 인자분석은 가정이 많은 제한적인 방법이므로 남용을 피해야 한다. 직교인자모형으로 그럴듯한 해석이 나오지 않으면 간단한 구조가 나타날 때까지 축을 회전(rotation)시킬 수 있다. 회전 방식으로는 varimax를 주로 사용한다. 기타 비직교모형에서 사용할 수 있는 oblique rotation 기법이 있다.

주식시세 자료에 대하여 두 개의 인자로 공분산구조를 설명하는 경우를 보자.

Variable	Estimated factor loadings	
	F_1	F_2
Allied Chemical	.783	-.217
du Pont	.773	-.458
Union Carbide	.794	-.234
Exxon	.713	+.472
Texaco	.712	+.524
cumulative explained proportion	.571	.733

결과 해석은 주성분분석의 경우와 마찬가지이다. Factor loading을 추정하는데는 주성분 방법과 최우추정법(MLE)의 두 가지가 있다. 다음은 최우추정법과 varimax 회전 최우추정의 결과이다.

	MLE		Rotated MLE	
Variable	F_1	F_2	F_1	F_2
Allied Chemical	.684	+.189	.601	.377
du Pont	.694	+.517	.850	.164
Union Carbide	.681	+.248	.643	.335
Exxon	.621	−.073	.365	.507
Texaco	.792	−.442	.208	.883
cumulative explained proportion	.485	.598	.335	.598

최우추정법에 의한 결과 해석은 전과 같다. 그러나 회전시킨 결과는 다르다. 회전시킨 후에 첫번째 인자는 화학주식에 공통된 경제적 상황이고 두번째 인자는 정유업에 고유한 경제적 상황을 반영한다. 앞서 언급한 시장 인자는 회전으로 파괴되었다. 회전 인자분석의 결과는 주성분분석이나 최우추정의 두번째 인자가 나뉘어진 것이다. 이런 경우에 회전은 전혀 가치가 없다. 왜냐하면 하나의 인자로 설명 가능한 것을 두 개의 인자로 설명하고 있기 때문에.

인자분석을 할 때는 공분산행렬을 기초로 수행할 것인가 상관행렬을 기초로 수행할 것인가, 주성분방법과 최우추정 중 어느 것을 사용할 것인가, 회전을 시켜야하는가 말아야 하는가, 직교모형을 사용할 것인가 비직교모형을 사용할 것인가 등 사용자가 고려해야할

요소가 많다. 가능하면 여러 방법을 사용해보고 가장 그럴듯한 해석이 나오는 분석을 찾을 수 밖에 없는데, 이런 이유로 작위적인 분석이라는 비판이 끊이지 않는다. 단, 전체를 지배하는 패턴이 아주 명확하다면 어떤 기법을 써도 결과가 크게 달라지지 않는다.

(3) 정준상관분석

이 분석에서는 두 그룹의 변수들 간의 연관도를 가장 잘 나타낼 수 있는 상관관계를 탐색한다. 즉, 한 그룹 내의 변수들의 선형조합과 다른 그룹 내의 변수들의 선형조합 중 가장 상관관계가 큰 선형조합들을 찾는다.

다음은 직업 만족도와 직업 특성 간의 관련성에 대한 정준상관분석이다. 어느 큰 회사의 784 명의 관리직 사원들을 대상으로 설문조사를 통하여 직업 특성에 대한 5 개 문항, 직업 만족도에 대한 7 개 문항을 조사했다. 직업 특성 변수와 직업 만족도 변수는 다음과 같다.

Job characteristic variables	Job satisfaction variables
feedback	supervisor satisfaction
task significance	career-future satisfaction
task variety	financial satisfaction
task identity	workload satisfaction
autonomy	company identification
	kind-of-work satisfaction
	general satisfaction

정준상관분석 결과, 가장 상관도가 높은 정준변수의 쌍은 다음과 같다.

$$U_1 = .42c_1 + .21c_2 + .17c_3 - .02c_4 + .44c_5$$

$$V_1 = .42s_1 + .22s_2 - .03s_3 + .01s_4 + .29s_5 + .52s_6 - .12s_7$$

이 두 정준변수 간의 상관도는 0.55이다. 직업 특성에서는 task identity의 기여도가 낮고, 직업 만족도는 주로 supervisor satisfaction, kind-of-work satisfaction, career-future satisfaction, company identification에 좌우된다.

(4) 분류

판별분석은 개체를 이미 알려진 그룹에 배정하는 기법이다. 예를 들어 인간은 남자와 여자의 두 그룹으로 나누어질 수 있다. 한 사람이 커튼 뒤에 있다. 이 사람이 남자인지

여자인지 어떻게 알 수 있을까? 판별 분석의 아이디어는 기존 남녀 데이타를 근거로 남녀를 구분하는 몇 가지 특성을 잰 다음, 이 특성을 근거로 판별 함수를 구축하고, 새로운 사람의 특성을 판별 함수에 넣어 남녀를 구분케하는 것이다. 한국 Gallup에서 지난 대통령 선거 결과를 정확히 예측할 수 있었던 근본 이유가 판별분석의 활용에 있다.

군집분석은 개체들을 몇 개의 그룹으로 나눌지 미리 알지 못하는 경우에 사용한다.

이 부분은 데이타가 커야하고 실례를 간단히 보일 방법이 없기 때문에 생략한다.

식품개발연구원의 쌀 분류 기법의 소개.

7. 반응표면분석(RSM)

Response Surface Methodology
RSM = 실험계획 + 회귀분석

반응표면분석은 조절가능한 입력변수(요인)들과 측정된 반응변수 간의 관계를 탐색하는 목적으로 사용되는 기법이다. 화학공업 분야에서 가장 많이 사용된다.

분석 단계:

① 적합시킬 모형의 선택과 적절한 실험설계
② 자료에 모형을 적합하고 적합도를 판정. 모형을 교정하고 필요하면 실험을 확장
③ 적합식을 이용한 예측과 요인들의 최적 조건 결정

<SAS 활용>
분석: RSREG, 삼차원도: G3D, 등고선도: GCONTOUR

<u>기초 개념</u>

η - 반응(response)
$\xi_1, \xi_2, \cdots, \xi_k$ - k 개의 입력변수
$x_1, x_2, \cdots, x_k$ - 부호화 변수(coded variable)

일반적으로 $\eta = f(\xi_1, \xi_2, \cdots, \xi_k)$ 이다.

그러나 함수 f의 형태는 알지 못한다.

RSM의 가치는 복잡한 형태의 관계 f를 관심 영역(region of interest) 내에서 저차의 다항식으로 근사시키는데 있다.

2차 모형(second order model) - 가장 많이 사용됨

$$\eta = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_i \sum_{\substack{j \\ i<j}} \beta_{ij} x_i x_j + \varepsilon$$

<u>입력 변수의 부호화(coding)</u>

입력변수들의 수준값을 −1에서 +1 사이로 변환하는 방법. 일종의 표준화.

단위에 의존하지 않는 분석 결과를 얻게된다.

예) 온도 요인: 100℃, 200℃
부호화: −1 , +1

예) 온도 요인: 100℃, 200℃, 300℃
부호화: −1 , 0 , +1

※ SAS에서는 자동적으로 부호화를 시행한다. 단, 중심합성계획 등 일부 실험계획에서는 부호화의 형태가 다를 수 있다.

최적점(optimum point)

최적점은 대체로 최대반응(maximum response), 또는 최소반응(minimum response)을 얻을 수 있는 요인들의 수준값이다.

적합된 2차 모형

$$\hat{\eta} = b_0 + \sum_{i=1}^{k} b_i x_i + \sum_{i=1}^{k} b_{ii} x_i^2 + \sum_{i} \sum_{\substack{j \\ i<j}} b_{ij} x_i x_j$$

미적분학에서 최대나 최소를 구할 때 함수를 미분하여 0으로 놓고 푸는 것을 상기하라. 반응표면 모형에는 변수가 k 개 존재하므로 적합식을 각 변수로 편미분한 뒤 0으로 놓고 풀어야 한다. 이 k 개 연립방정식의 해를 정상점(stationary point)이라 부른다.

정상점은 경우에 따라 최대점, 최소점, 안장점(saddle point)이 될 수 있다.

정준형식(canonical form)

정상점을 원점으로 하고 등고선의 축을 주축으로 간주하는 좌표축 변환. 해석이 쉬워진다.

$$\hat{y} = \hat{y}_0 + \lambda_1 w_1^2 + \lambda_2 w_2^2 + \cdots + \lambda_k w_k^2$$

$\hat{y}_0$ − 정상점에서 추정된 반응값

$w_1, w_2, \cdots, w_k$ − 새로운 좌표축 변수

$\lambda_1, \lambda_2, \cdots, \lambda_k$ − 계수, 고유값(eigenvalue)

고유값 $\lambda_1, \lambda_2, \cdots, \lambda_k$들이	정상점은
모두 음수이면	최대점
모두 양수이면	최소점
음수, 양수가 섞여있으면	안장점
0이 존재하면	능선체계

<u>좌표축 변환의 도해</u>

칠판에 그릴 것.

예) $\hat{y} = 10 - 5w_1^2 - 2w_2^2$, 정상점 $= (0,0)$

정상점은 최대점. 요인 w_1이 요인 w_2보다 반응에 끼치는 영향이 크다.

※ 때때로 정상점이 실험영역 밖에 존재할 수 있다. 이런 경우에는 능선분석(ridge analysis)을 실시한다. 능선분석은 정상점이 안장점일 때도 해야한다.

예) 어떤 화학 물질을 **최대**로 생산할 수 있는 처리 시간과 온도를 찾기 위하여, 몇 가지 처리 시간과 온도의 수준 조합점들에서 실험을 하였고, 각 실험점에서 생산량을 측정하였다.

time	temp	mbt
4.0	250	83.8
20.0	250	81.7
12.0	250	82.4
12.0	250	82.9
12.0	220	84.7
12.0	280	57.9
12.0	250	81.2
6.3	229	81.3
6.3	271	83.1
17.7	229	85.3
17.7	271	72.7
4.0	250	82.0

출력 상단에 요인 변수들의 부호화 과정이 나와있다. 즉, 요인 변수 time에 대하여는 각 데이타 값들에서 12를 빼고 8로 나누어 부호화를 하였으며, temp는 250을 빼고 30으로 나누어 부호화를 하였다.

전체 모형에 대한 유의 수준은 0.05보다 작으므로 가정된 모형은 자료에 잘 적합된다고 할 수 있다.

적합된 반응표면 모형은

$$\hat{y} = -545.87 + 6.87(time) + 4.99(temp)$$
$$+0.022(time)^2 - 0.0098(temp)^2 - 0.030(time \times temp)$$

부호화된 요인 변수 time'과 temp'로 표기하면

$$\hat{y} = 82.17 - 1.01(time') - 8.68(temp')$$
$$+1.38(time')^2 - 8.85(temp')^2 - 7.22(time' \times temp')$$

정상점의 좌표: 부호화된 자료를 기준으로 하면 (time',temp')=(−0.44,−0.31)

정상점의 좌표: 원래 자료를 기준으로 하면 (time,temp)=(8.47,240.70)

정상점에 대한 반응값은 83.74이다.

적합된 반응 모형을 정준 형식으로 바꾸면,

$$\hat{y} = 83.74 + 2.53w_1^2 - 10.00w_2^2$$

두 고유값의 부호가 서로 다르므로 정상점은 안장점이다. time에 대한 고유값이 2.53으로 양수이고, temp에 대한 고유값은 −10.00으로 음수이며, temp에 대한 고유값이 절대값으로는 time에 대한 고유값보가 크다. 우리가 현재 찾으려는 최적해는 최대점이므로, 처리 시간을 길게하고 온도 수준을 낮게 잡으면 최대 생산이 될 것이다.

능선분석 결과를 보면, 부호화된 원점으로부터 반경 1의 원주 상의 점들 중, 처리 시간이 18.45이고 온도가 232.26일 때 최대 생산량 87.73을 얻게됨을 알 수 있다.

<u>기타:</u>

결정계수의 값은 0.8. 적합이 잘 되고 있다. 교차곱에 대한 제곱합은 51.84로 모형으로 설명 가능한 변동 512.19의 약 10% 정도이다. 따라서 두 요인 간의 상호작용은 무시할 수 있다.

SAS

Coding Coefficients for the Independent Variables

Factor	Subtracted off	Divided by
TIME	12.000000	8.000000
TEMP	250.000000	30.000000

Response Surface for Variable MBT

Response Mean	79.916667
Root MSE	4.615964
R-Square	0.8003
Coef. of Variation	5.7760
Sum of Squared Residuals	127.842720
Predicted Resid SS (PRESS)	791.336033

Regression	Degrees of Freedom	Type I Sum of Squares	R-Square	F-Ratio	Prob > F
Linear	2	313.585803	0.4899	7.359	0.0243
Quadratic	2	146.768144	0.2293	3.444	0.1009
Crossproduct	1	51.840000	0.0810	2.433	0.1698
Total Regress	5	512.193947	0.8003	4.808	0.0410

Residual	Degrees of Freedom	Sum of Squares	Mean Square
Total Error	6	127.842720	21.307120

SAS

Parameter	Degrees of Freedom	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEPT	1	-545.867976	277.145373	-1.970	0.0964
TIME	1	6.872863	5.004928	1.373	0.2188
TEMP	1	4.989743	2.165839	2.304	0.0608
TIME*TIME	1	0.021631	0.056784	0.381	0.7164
TEMP*TIME	1	-0.030075	0.019281	-1.560	0.1698
TEMP*TEMP	1	-0.009836	0.004304	-2.285	0.0623

Parameter	Parameter Estimate from Coded Data
INTERCEPT	82.173110
TIME	-1.014287
TEMP	-8.676768
TIME*TIME	1.384394
TEMP*TIME	-7.218045
TEMP*TEMP	-8.852519

Factor	Degrees of Freedom	Sum of Squares	Mean Square	F-Ratio	Prob > F
TIME	3	61.290957	20.430319	0.959	0.4704
TEMP	3	461.250925	153.750308	7.216	0.0205

Canonical Analysis of Response Surface
(based on coded data)

Factor	Critical Value Coded	Critical Value Uncoded
TIME	-0.441758	8.465935
TEMP	-0.309976	240.700718

Predicted value at stationary point 83.741940

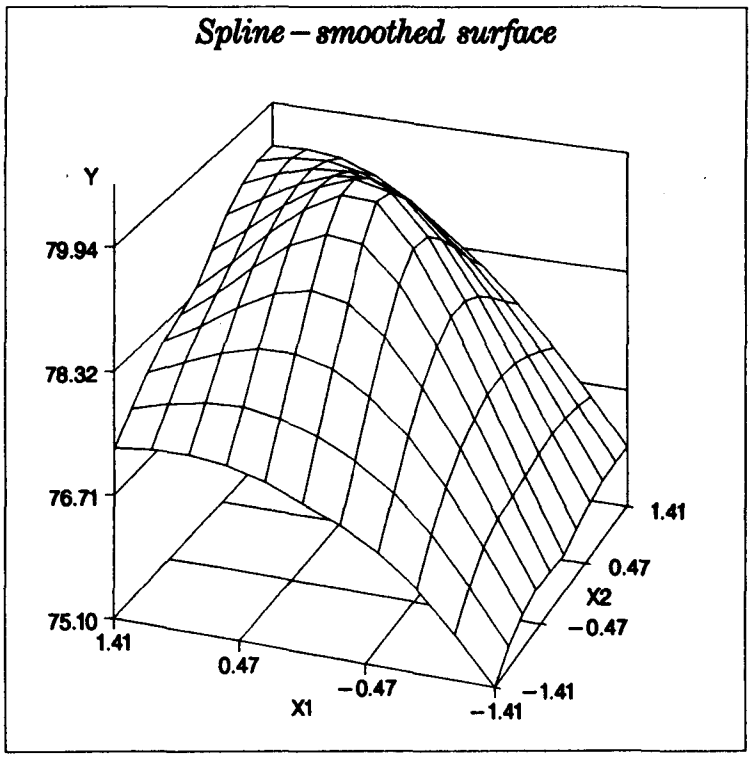
Eigenvalues	Eigenvectors TIME	Eigenvectors TEMP
2.528816	0.953223	-0.302267
-9.996940	0.302267	0.953223

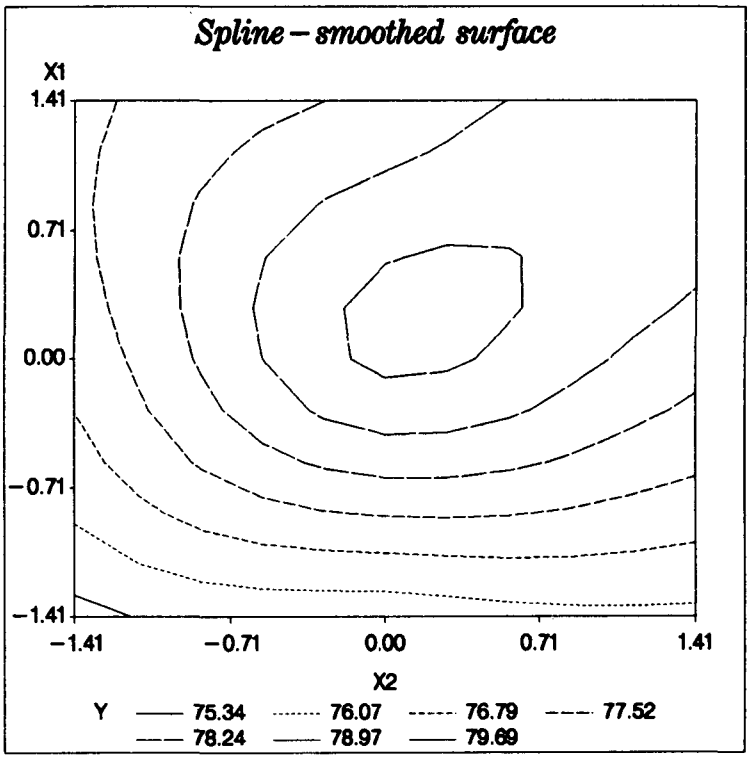
Stationary point is a saddle point.

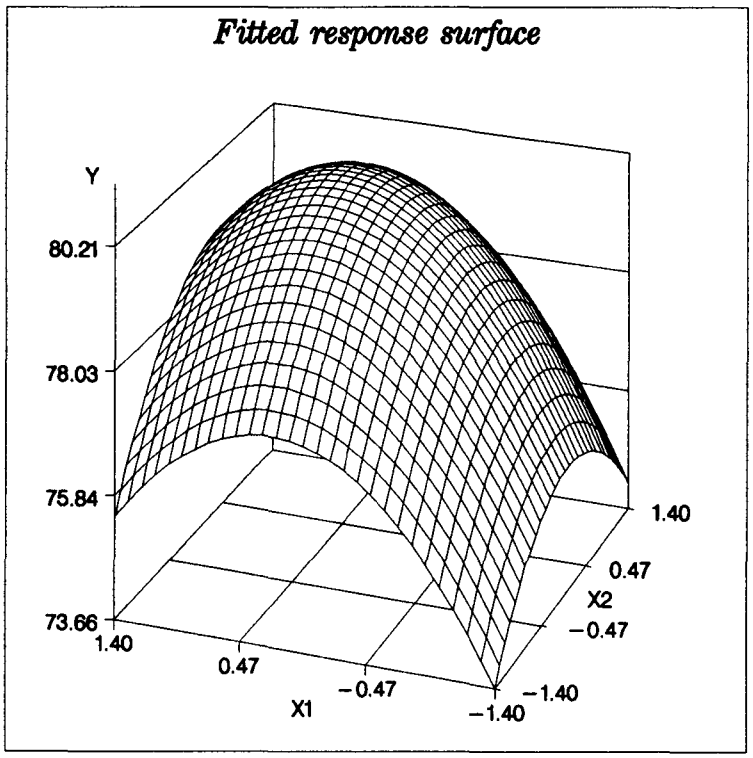
Estimated Ridge of Maximum Response for Variable MBT

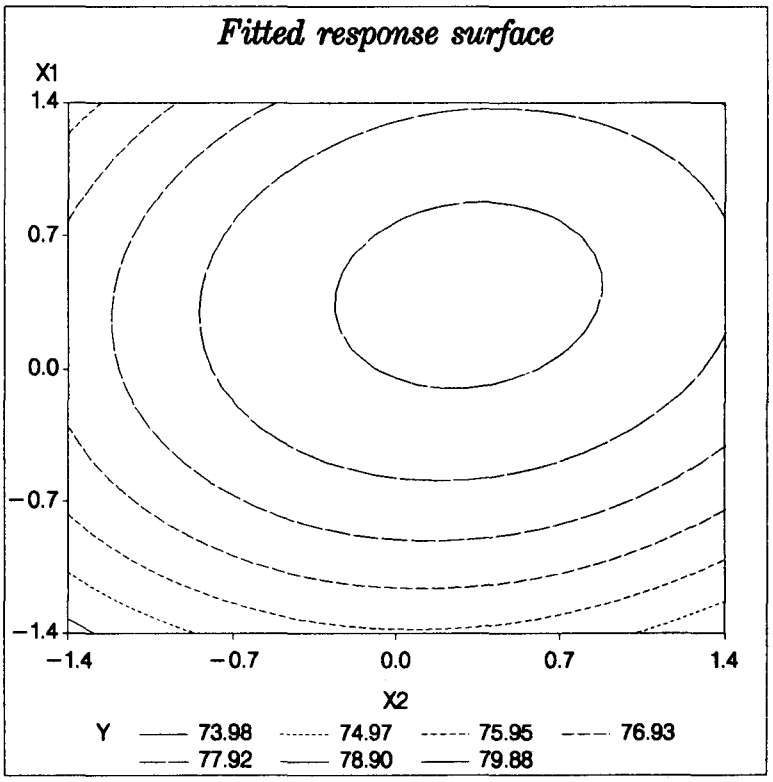
Coded Radius	Estimated Response	Standard Error	Uncoded Factor Values TIME	Uncoded Factor Values TEMP
0.0	82.173110	2.665023	12.000000	250.000000
0.1	82.952909	2.648671	11.964493	247.002956
0.2	83.558260	2.602270	12.142790	244.023941
0.3	84.037098	2.533296	12.704153	241.396084
0.4	84.470454	2.457836	13.517555	239.435227
0.5	84.914099	2.404616	14.370977	237.919138
0.6	85.390012	2.410981	15.212247	236.624811
0.7	85.906767	2.516619	16.037822	235.449230
0.8	86.468277	2.752355	16.850813	234.344204
0.9	87.076587	3.130961	17.654321	233.284652
1.0	87.732874	3.648568	18.450682	232.256238

반응표면분석: 3차원 표면도와 등고선도의 실례




Spline – smoothed surface
X1
1.41
0.71
0.00
−0.71
−1.41
−1.41
−0.71
0.00
0.71
1.41
X2
Y
75.34
76.07
76.79
77.52
78.24
78.97
79.69


Fitted response surface
Y
80.21
78.03
75.84
73.66
1.40
0.47
−0.47
−1.40
X1
1.40
0.47
X2
−0.47
−1.40


Fitted response surface
X1
1.4
0.7
0.0
−0.7
−1.4
−1.4
−0.7
0.0
0.7
1.4
X2
Y
73.98
74.97
75.95
76.93
77.92
78.90
79.88