

Kwon, Wolfpack Sehyug

Linear Model (Prediction Model)



1

Introduction

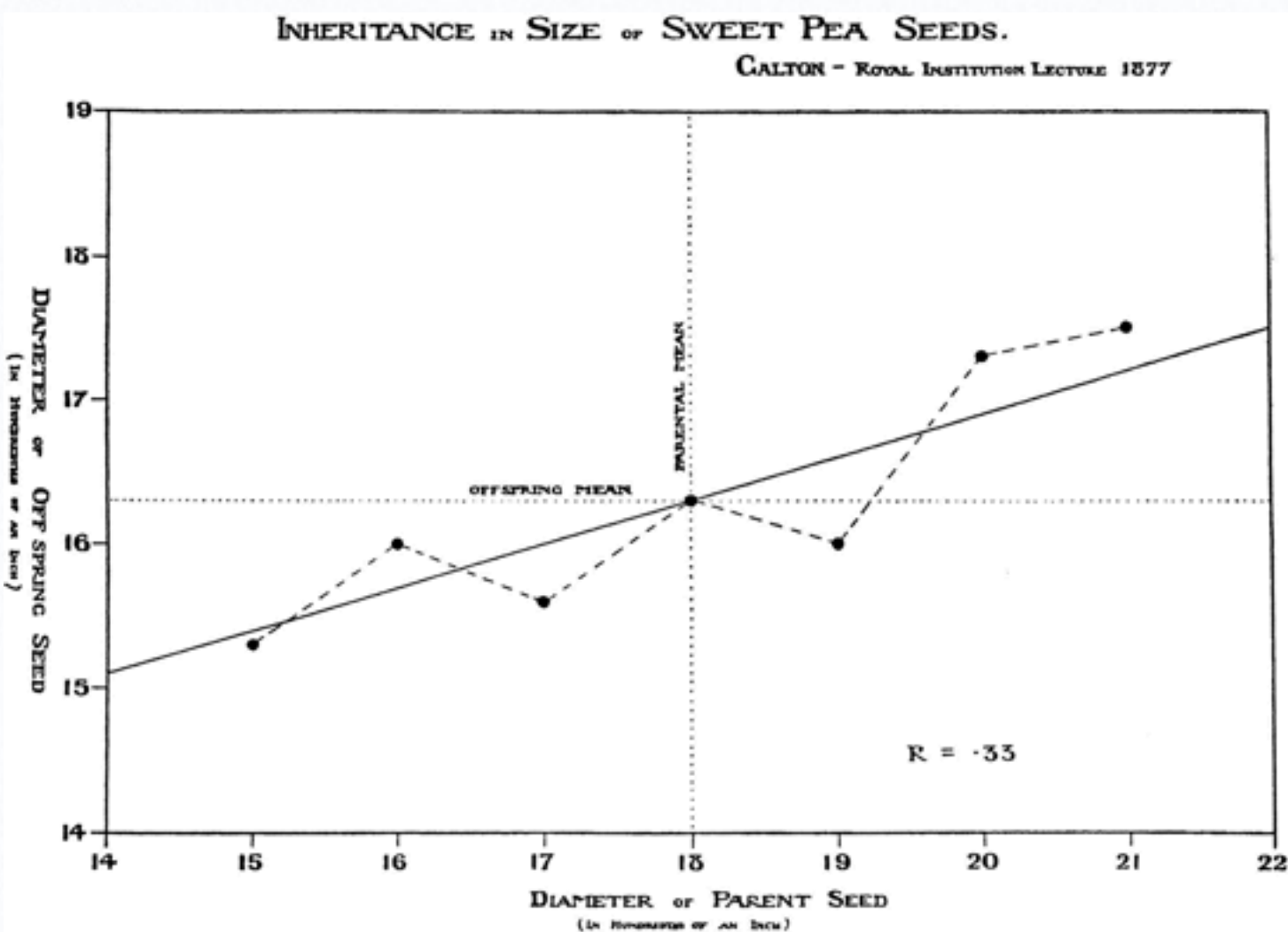
Creativity is just connecting things.

–Steve Jobs

History

Regression

- 회귀(regress)의 사전적 의미 “go back to an earlier and worse condition” (옛날 상태로 돌아감)
- 영국의 유전학자 Francis Galton(1822-1911)의 연구 -(처음에는 sweat pea)



부모의 키와 자녀의 키 사이 관계 연구 : 928명의 성인 자녀 키(여자는 키에 1.08배, 행)와 부모 키(아버지 키와 어머니 키의 평균, 열)를 조사하여 다음 표를 얻었음

Galton은 표를 관찰한 결과 : 키는 평균(아래 그림 M) 수준을 기준으로 부모와 자식의 편차가 커지고 있음을 알 수 있음 - 부모의 키가 평균 키 이상이면 자녀의 키는 부모 키보다 작아지는 경향을 보이고 있고 평균 키 이하에서는 반대 현상 표에서 직선이 중앙 일정 구간이외에는 중심으로 향해 낮아짐

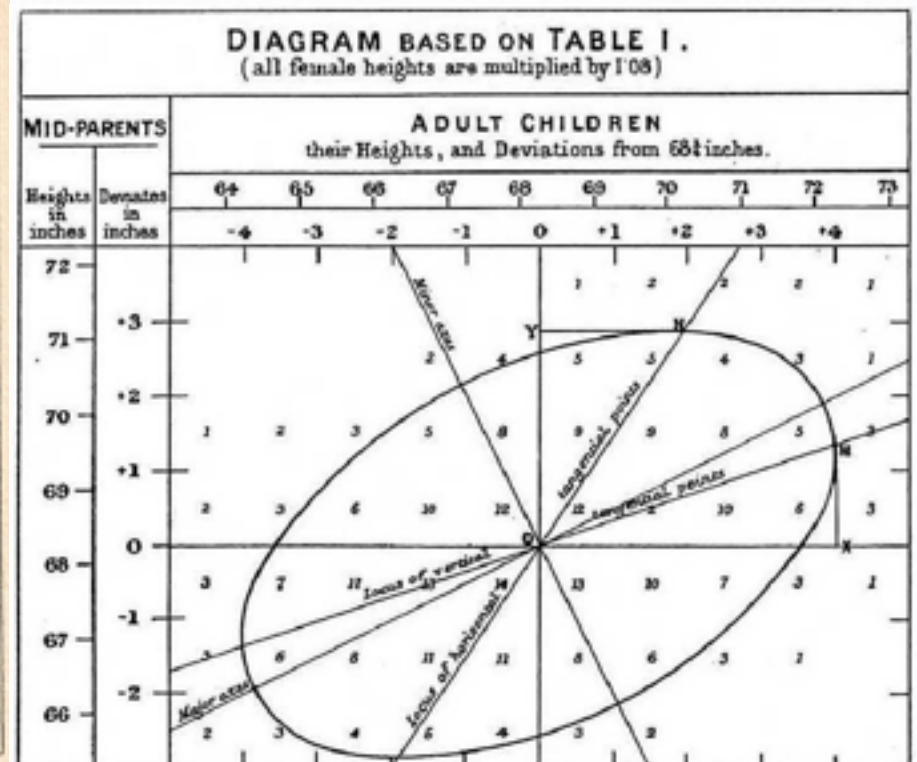
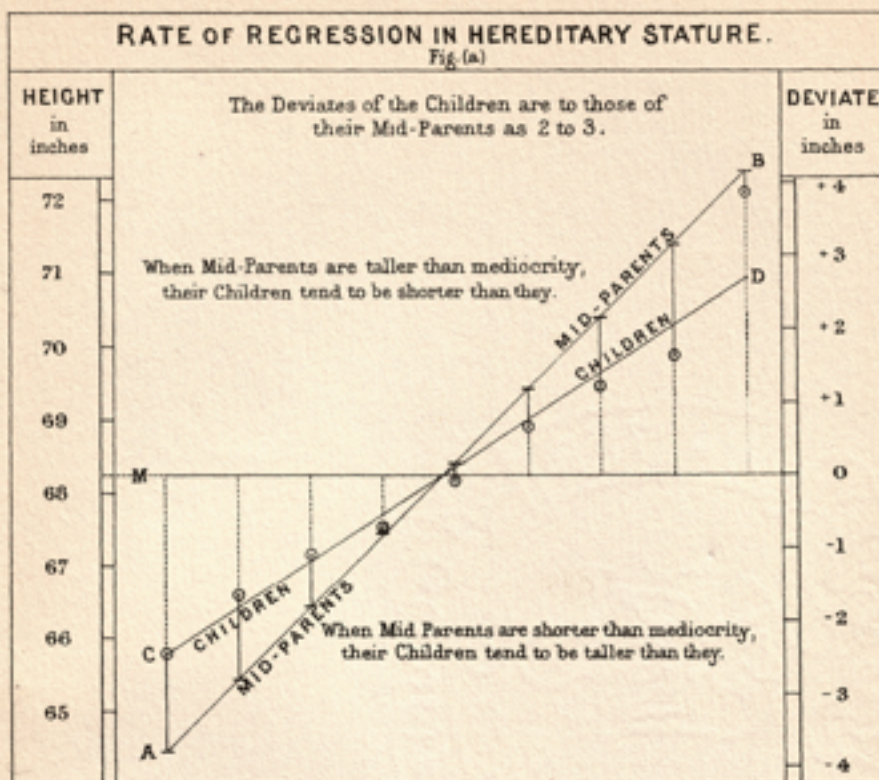
즉 중심으로 회귀하려는 경향, 그가 제안한 분석방법의 이름을 “회귀”분석이라 명명하였다.

NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.

(All Female heights have been multiplied by 1.08).

Heights of the Mid-parents in inches.	Heights of the Adult Children.														Total Number of		Medians.
	Below	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72.5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72.2
71.5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69.9
70.5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69.5
69.5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68.9
68.5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68.2
67.5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67.6
66.5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67.2
65.5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66.7
64.5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65.8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66.3	67.8	67.9	67.7	67.9	68.3	68.5	69.0	69.0	70.0	..	..	..	..	..

Plate IX.



Karl Pearson(1903) :

Galton의 아이디어 수리적 모형화 및 추정방법 OLS(ordinary least square 최소자승법) 제안하고, 1,078명 부자 키 데이터 활용하여 $Son_H = 33.73 + 0.516Father_H$ 추정식 얻음

- 아버지의 키가 1인치 커지면 아들 키는 0.516인치 커짐 - 기울기 개념

분산분석

- 두 모집단 평균 검정을 위하여 사용되던 z-검정 방법, w.s. Gosset에 제안한 소표본 검정방법인 t-분포 사용되었음
- Fisher의 저서 "Statistical Methods for Research Workers." : k 개 정규 모집단으로 평균 비교를 확장한 개념인 ANOVA 제안
- 분산 분석 (ANOVA)은 데이터의 총 변동성을 체계적 요인과 랜덤 요인으로 나눈 통계에 사용되는 분석 도구입니다.
- $F = \frac{MST}{MSE}$: 변동 비에 의해 체계적 변동의 유의성을 검정

회귀분석과 분산분석의 같은 점 및 다른점 $Y = f(X's) + e$

같은 점

- 목표변수 Y 를 가지고 있고 정규분포를 따른다. (사실 오차항이 정규분포를 따르고 X 는 확률변수가 아닌 결정변수이므로 목표변수가 정규분포를 따른다)
- 예측변수(X , Input)과 목표변수(output) 함수관계 $f(x's)$ 는 선형이다. => 두 분석을 합하여 선형모형 분석이라 한다.

다른점

예측변수의 형태

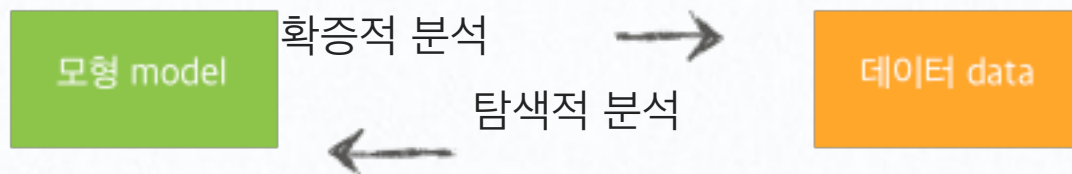
- 회귀모형의 예측변수는 측정형 - 범주형을 사용할 때는 지시변수 형태를 따름
- 분산분석의 예측변수는 요인(처리효과)라 하며 범주형임 - 순서형인 경우 RSM response surface model, 측정형인 경우는 ANCOVA 모형이라(공변량 covariate) 함

유의성 검정

- 회귀모형 : 예측변수의 기울기 유의성 - 회귀모형 전체의 유의성은 모형변동, 오차변동 비 검정
- 분산분석 : 요인변동, 오차변동 비

Data

모형과 데이터 사이클



과학에서 이론이 제안되고 데이터 분석이 이루어지는 경우보다는 데이터로부터 새로운 이론이나 모형을 도출하는 경우가 많고 탐색적 자료 분석에 의해 제안된 이론이나 모형은 다시 confirmatory 방법에 의해 유의성이 (significance) 검증되므로 모형과 데이터는 순환 사이클을 갖는다.

통계적 모형은 과학적 진실이기 보다는 분석 대상이 되는 사실(현황)의 대표적 모형이다. 예를 들어, 회귀 모형에서는 $(Y = \alpha + \beta X + e)$ 선형함수(모형)이 설명하지 못하는 오차항(e)이 존재하고 이 오차항은 평균 0, 분산 σ^2 인 정규분포를 따른다고 가정한다.

데이터 정의

(정의) 추론, 정보획득, 계산에 사용되는 실제의 조사, 측정된 값 (통계) (웹스터), 정보를 가진 숫자의 모임

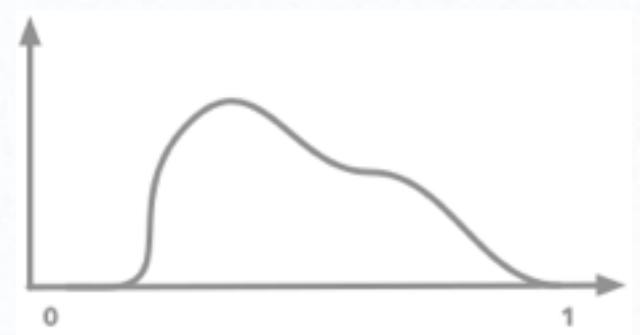
통계학에서 분석 대상인 데이터는 행과 열로 이루어져 있으며, 행은 개체 subject 관측치, 열은 관심 특성, 변수 variable로 이루어진 숫자 행렬이다. 행의 첨자는 개체를 나타내고, 열의 첨자는 변수를 나타낸다.

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

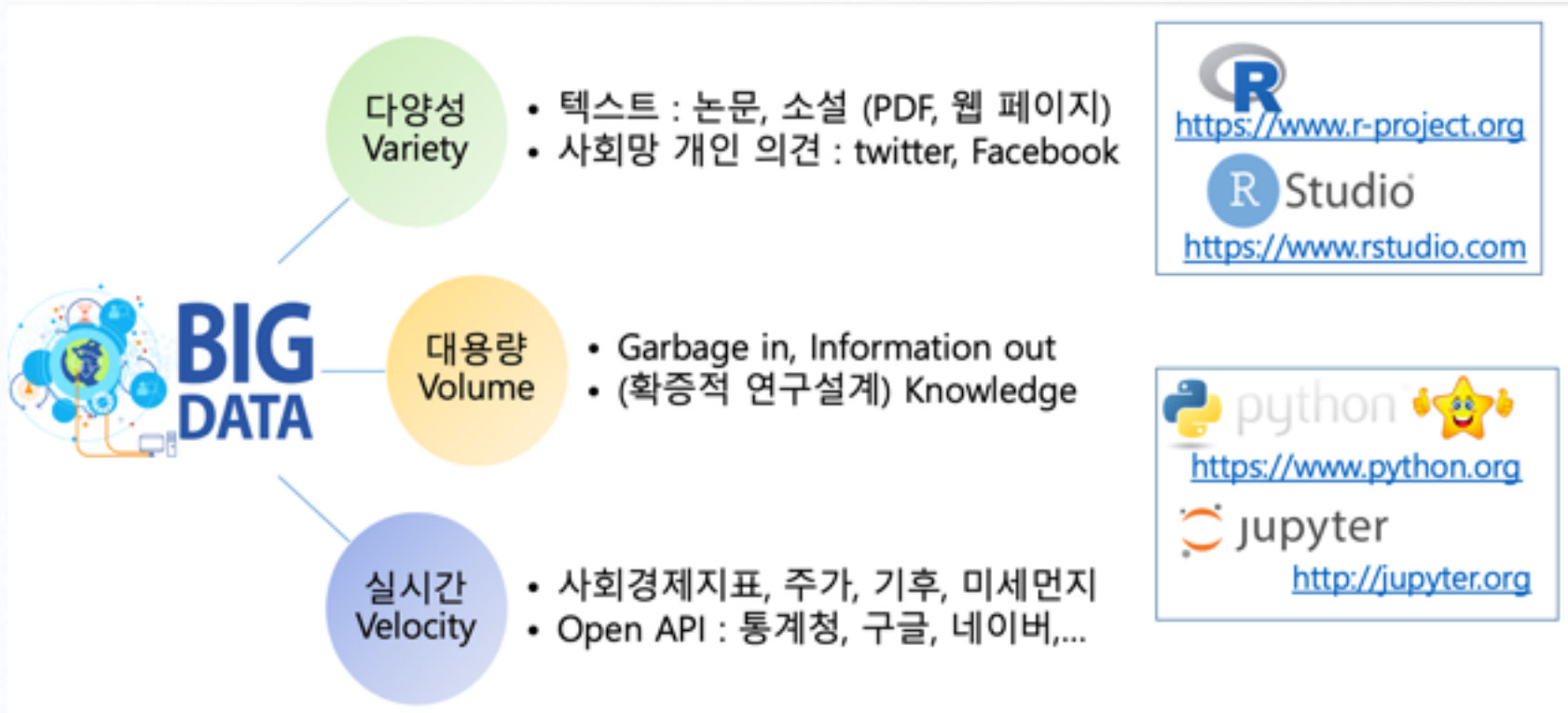
데이터와 확률변수 random variable

통계분석의 대상이 되는 데이터는 변수들로 구성되어 있는데, 변수의 관측결과와 숫자를 일대일 매칭한 함수를 확률변수라 한다. 변수의 관측치가 숫자인 경우는 데이터 값 자체가 확률변수이다.

확률변수가 가질 수 있는 값과 대응하는 확률을 일대일 매칭한 함수를 확률분포함수(pdf probability density function)라 한다. PDF 는 데이터(변수)가 가진 모든 정보를 표현하고 있다. 변



수의 관측값이 가질 수 있는 가장 큰 값, 작은 값, 중앙 위치, 그리고 어느 구간이 관측 가능성이 가장 높은 지 등, 모든 정보를 확률밀도함수가 가지고 있다



데이터 종류

고전적 정의 : 통계학에서는 변수 **variable**, 사회과학방법론에서는 **scale**라 한다.

질적 Qualitative, Non-metric (분류형 변수, classified, 범주형 categorical) : 개체를 분류하기 위해 측정된 변수를 의미 하며 성별, 결혼여부 등이 그 예이다.

- 명목 (nominal): 개체를 구분하거나 분류할 목적 (예) 성별, 결혼여부, 직업, 거주시도
- 순서, 서열 (ordinal): 개체를 일정한 순위순서를 가진 분류, 성적(A, B, ..) 소득수준(상, 중, 하),

양적 Quantative, Metric (측정형 변수, measurable) : 실험 개체의 측 정 가능한 특성을 측정한 변수로 키, 몸무게, 평점, IQ, 교통량, 사망자 수가 그 예이다. 연속형 변수는 모두 측정형 변수이고 이산형 변수 중 측정형 변수가 있을 수 있다.

- 구간, 등간 interval : 0 의미 없고 배율도 의미 없음 (예) 온도, 리커트 척도
- 비율척도 ratio : 0이라는 숫자의 의미가 있고 배율의 의미 존재 (예) 모든 측정형 소득

시간적 정의

- 횡단 변수 cross section : 일정 시점의 조사 데이터
- 종단변수 time series : 시간적 순서를 갖는 데이터 - 대신 로 표현, 경제 지표(환율, 수출량)나 기업의 연차별 자료(연도별 매 출액), 연도별/월별 청년 실업율 - 계량 경제(econometrics), 패널회귀분석

인과관계 casual relationship

인과 관계 모형은 통계분석에 의해 검증되는 것이지 발견하는 것은 아니다. 모형 설정은 이론, 경험적 타당성에 근거하여 이루어진다.

- Input X - 독립변수, 요인(처리효과), 예측변수, 예측변수, 내생(endogenous)변수 : 원인이 되는 변수, 분산 분석에서 설명 변수는 처리 효과, 요인으로 불리어진다.
- Output Y - 목표변수, 반응변수, 목표변수, 외생(exogenous)변수 : 영향을 받는 변수

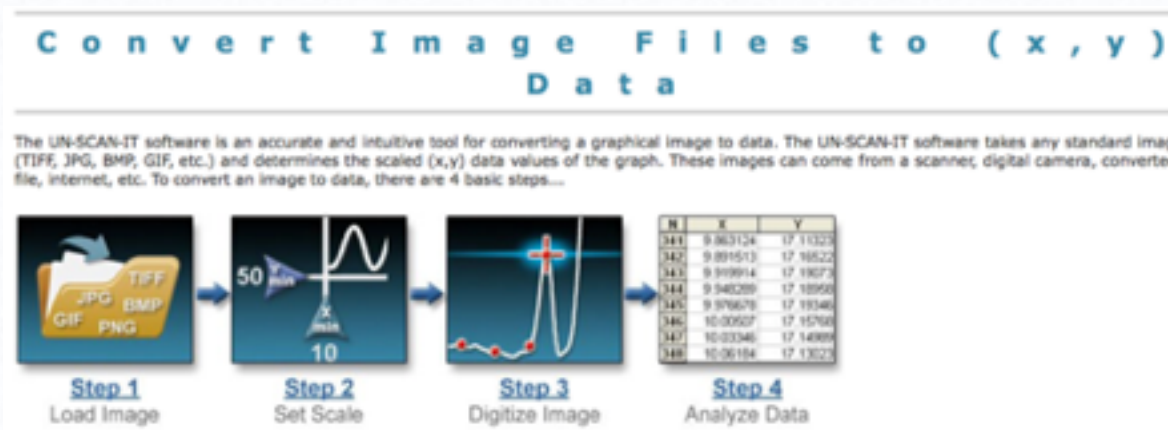
일반적 통계 모형 - $Y = f(X's) + e$: 함수 f 형태 중 해석이 용이하고 간단한 모형이 선형

인과 관계는 이론적, 경험적 타당성에 근거하여 데이터 수집 전에 이론이나 경험에 의해 설정되는 것이지 분석 후 설정되는 것은 아님

- 예를 들어 보자. 대학생 40명의 GPA성적과 용돈을 조사하였다. GPA성적-Y, 용돈을 X로 하여 회귀 분석을 실시한 결과 유의한(significant)가 있다고 해서 용돈이 수능성적에 영향을 미친다고 할 수 있나? No, 인과관계 설명은 또 다른 이야기임, 최종 인과관계는 사회실험 조사 필요
- 고전적인 예제 - 흡연과 암의 인과관계 : 횡단시점 분석(코호트 분석), 암 환자의 흡연비율과 일반인 흡연비율 차이 비교 - 그러나 숨은 변인 효과
- association (연관)이 인과 (causation)를 의미하는 것은 아님 - 두 변수의 관계를 분석할 때 고려하지 않았으나 관계에 영향을 미치는 변수를 lurking(숨은, 잠재) 변수로 인한 것임, 흡연과 암의 관계에서 소득수준이 잠재변인일 수 있음, 소득수준 효과를 제외하면 흡연과 암의 (인과)관계가 존재하지 않을 수 있음

비정형 포맷

- 문자 : 텍스트 마이닝, 자연어 처리, word cloud
- 음성 : 오디오 파일 포맷을 데이터로 변환/역변환 가능 (convert audio format into data format)
- 이미지 : 이미지 파일 (x, y) 좌표나 6자리 숫자 단위로 변환 가능



데이터 포맷 in Big Data

Data frame

- 열이름과 행 이름을 가진 행렬 데이터
- 열 이름은 변수명과 일치하며, 행은 일련의 숫자로 되어 있음
- 모든 연산은 변수명(열 단위), 행 단위로 하게 된다. => 함수를 이용한다.

Matrix, Array

- 행과 열로 이루어진 행렬과 열 하나만 이루어진 벡터 형태인 array
- 숫자로 구성되어 있고 연산은 수학 (행렬) 연산자에 의해서만 가능하다.

언어에서의 구별

내용	R (typeof())	Python
데이터	dataframe	Pandas : pd.DataFrame, pd.Series
행렬	matrix(), vector()	Numpy : np.matrix, np.array
문자열	list(), c(" ")	str
숫자	c()	int, float, complex
순서(열거) sequence		list, tuple, range
Boolean	논리 연산자 &, &&	참, 거짓
Operators		+, -, *, /
		+=, %=, ==, !=,

matrix algebra

행렬

정의

$$A_{n \times p} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{np} \end{bmatrix} \Rightarrow A_{3 \times 3} = \begin{bmatrix} 1 & 2 & -1 \\ 0 & 1 & 2 \\ 1 & 0 & 2 \end{bmatrix}, B_{3 \times 2} = \begin{bmatrix} 1 & -1 \\ 2 & 0 \\ 1 & 2 \end{bmatrix}$$

- 간편식 $\{A_{ij}\}$ - 첨자 i는 행, j는 열을 나타냄
- 행 차수 n, 열 차수 p인 행렬, 차수 (nXp)인 행렬 - 기호 대문자 A, B, ...

벡터 vector

- 열의 차수가 1인 행렬을 열 벡터(column vector) $\Rightarrow \underline{x} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$
- 행의 차수가 1인 행렬을 행 벡터(row vector) $\Rightarrow \underline{y} = [1 \ 2 \ 1]$
- 일반적으로 벡터라 함은 열 벡터를 의미하고 소문자와 아래 바를 사용 - 기호 $\underline{a}, \underline{b}, \dots$

스칼라 scalar

- 행과 열의 차수 모두가 1인 경우 행렬은 scalar (스칼라)이다.
- 즉 -2, 3 등의 실수는 행렬에서는 스칼라라 한다. - 기호 a, b, ...

특수한 행렬

정방 square 행렬

- 행과 열의 차수가 동일한 행렬 - $A_{n \times n}$

대각 diagonal 행렬

- 정방행렬 중 대각원소를 제외한 모든 원소가 0인 경우 - $D_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -2 \end{bmatrix}$
- 대각원소의 합을 trace라 한다. $tr(D) = \sum_{i=1}^n d_{ii} = -1$, $tr(A) = 16$

항등 identity 행렬

- 대각행렬 중 대각원소를 모두 1인 경우 : $I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$
- 항등행렬의 trace는 차수 n 이고 숫자의 1과 같은 역할을 한다.

전치 transpose

- 행렬 $A_{n \times p}$ 에서 행과 열을 바꾸는 것을 전치라 하고 A' (혹은 A^T)로 나타내며 차수는 $(p \times n)$.
- $A'_{3 \times 3} = \begin{bmatrix} 1 & 0 & 1 \\ 2 & 1 & 0 \\ -1 & 2 & 2 \end{bmatrix}$, $\underline{x}' = [1 \quad 2 \quad 1]$

대칭 symetric 행렬

- 원 행렬과 대칭 행렬이 동일한 행렬 : $A' = A$ (예) $S_2 = \begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}$ 는 대칭행렬
- 항등행렬은 모두 대칭 행렬이다.

사칙연산 Algebra

행렬의 합(차) addition / subtraction

- 대응하는 원소의 합/차로 계산된다.
- 합(차) 계산이 가능하려면 두 행렬의 차수는 동일해야 한다. conformable

(예제) $A \pm B$ 는 성립하지 않는다.

$$B_{3 \times 2} = \begin{bmatrix} 1 & -1 \\ 2 & 0 \\ 1 & 2 \end{bmatrix}, C_{3 \times 2} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ -2 & -1 \end{bmatrix} \Rightarrow C_{3 \times 2} = A + B = \begin{bmatrix} 1 & 1 \\ 5 & 4 \\ -1 & 1 \end{bmatrix}$$

행렬의 곱 multiplication

- 행렬에서 곱의 conformable(적합) 조건 : 앞 행렬(벡터)의 열의 차수와 뒤 행렬 행의 차수가 동일
- 곱의 결과는 (앞 행렬의 행 차수)x(뒤 행렬의 열의 차수)인 행렬(벡터, 스칼라)이다.
- 행렬A[i번째 행]x행렬B[j번째 열] 대응 곱 =(AB)행렬[(i, j) 원소]

$$AB_{kj} = \sum_i^p a_{ki} b_{ij}$$

(예제) AB 는 계산 적합하나 AB' 는 계산 적합하지 않다.

$$A_{3 \times 3} = \begin{bmatrix} 1 & 2 & -1 \\ 0 & 1 & 2 \\ 1 & 0 & 2 \end{bmatrix}, B_{3 \times 2} = \begin{bmatrix} 1 & -1 \\ 2 & 0 \\ 1 & 2 \end{bmatrix} \Rightarrow C_{3 \times 2} = (AB)_{3 \times 2} = \begin{bmatrix} 4 & -3 \\ 4 & 4 \\ 2 & 3 \end{bmatrix}$$

(성질)

- $AB \neq BA$ (일반적으로)
- $(AB)' = B'A'$: 곱의 전치
- $A'A$ 은 대칭행렬이다.

(행 벡터)x(열 벡터)

벡터의 곱은 앞 벡터의 열의 원소와 대응하는 뒤 벡터의 행의 원소의 곱을 더한 값을 적으면 된다. 곱이 가능하기 위해서는 앞 행의 차수와 열의 차수는 같아야 하며 행 벡터(1xp행렬)와 열 벡터(px1행렬) 곱은 scalar(1x1행렬)이다.

$$\underline{c} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \underline{d}' = [1 \quad 1 \quad -1] \Rightarrow \underline{d}'\underline{c} = 0 \text{ 결과 값은 스칼라(scalar)}$$

(열 벡터)x(행 벡터)

열 벡터 열 원소와 행 벡터의 행 원소의 곱을 (앞의 열 벡터 원소 위치) 행, (뒤의 행 벡터 원소 위치) 열로 하여 행렬을 만든다. 앞의 열 벡터와 차수와 뒤 열 벡터 차수는 같을 필요가 없다. 결과는 (nxp) 행렬이다.

$$\underline{c} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \underline{d}' = [1 \quad 1 \quad -1] \Rightarrow \underline{c}'\underline{d} = \begin{bmatrix} 1 & 1 & -1 \\ 2 & 2 & -2 \\ 1 & 1 & -1 \end{bmatrix} \Rightarrow \text{결과 값은 행렬}$$

행렬식 Determinant

- 행렬식은 정방행렬에서만 계산되며 결과는 스칼라 (기호) $|A|$ 혹은 $\det(A)$

$$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = a \cdot d - c \cdot b$$

$$\det \begin{bmatrix} 8 & 3 \\ 4 & 2 \end{bmatrix} = 8 \cdot 2 - 4 \cdot 3 = 16 - 12 = 4 \quad (2 \times 2 \text{ 행렬})$$

(3 × 3 행렬)

소인자 Minor

- 지정한 행, 열을 제외한 행렬 :

$$M_{23} = \det \begin{bmatrix} 1 & 4 & \square \\ \square & \square & \square \\ -1 & 9 & \square \end{bmatrix} = \det \begin{bmatrix} 1 & 4 \\ -1 & 9 \end{bmatrix} = (9 - (-4)) = 13$$

- 소인자에 의한 행렬식 계산 :

$$|A| = \sum_{i=1}^n a_{ij}(-1)^{i+j} |M_{ij}| = \sum_{j=1}^n a_{ij}(-1)^{i+j} |M_{ij}|$$

차수가 3일 경우: $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 7 \\ 8 & 9 & 10 \end{bmatrix}$ 의 행렬식은

$$|A| = 1 * (-1)^{1+1} \begin{vmatrix} 5 & 7 \\ 9 & 10 \end{vmatrix} + 2 * (-1)^{1+2} \begin{vmatrix} 4 & 7 \\ 8 & 10 \end{vmatrix} + 3 * (-1)^{1+3} \begin{vmatrix} 4 & 5 \\ 8 & 9 \end{vmatrix} = 7 \text{ (1번째 행 이용)}$$

$$|A| = 4 * (-1)^{2+1} \begin{vmatrix} 2 & 3 \\ 9 & 10 \end{vmatrix} + 5 * (-1)^{2+2} \begin{vmatrix} 1 & 3 \\ 8 & 10 \end{vmatrix} + 7 * (-1)^{2+3} \begin{vmatrix} 1 & 2 \\ 8 & 9 \end{vmatrix} = 7 \text{ (2번째 행 이용)}$$

(행렬식 성질)

- $|A'| = |A|$, $|AB| = |BA|$, $|AB| = |A| |B|$
- 행렬 A의 두 행이 같으면 행렬식은 0이다.
- 한 행(열)의 상수를 곱하여 다른 행에 더해도 행렬식 값은 변하지 않는다.
- 한 행(열)을 다른 행들의 선형 결합으로 표현할 수 있으면 행렬식 0이다.

역행렬

- $AB = I$ 만족하는 행렬 B를 행렬 A의 역행렬이라 하고 A^{-1} 로 표기한다. $A^{-1} = \frac{1}{|A|} \text{adj}A = \frac{1}{|A|}$

(역행렬 성질)

- 역행렬은 unique하다.
- $|A^{-1}| = 1/|A|$, $(A^{-1})^{-1} = A$, $(A')^{-1} = (A^{-1})'$, $(AB)^{-1} = B^{-1}A^{-1}$

계수 rank

- (정의) 행렬에서 선형 독립인 행의 개수와 열의 개수 중 최소값, $\text{rank}(A_{m \times n}) \leq \min(n, m)$
- 정의(full rank) : $(n \times n)$ 정방 행렬에서 선형 독립인 행(열)의 개수()가 행렬의 차수 n와 같다면 이 행렬은 full-rank 행렬이라 한다. 즉 $\text{rank}(A_{n \times n}) = n$ 이면 full-rank이다.

정의(LIN: linearly independent vector)

- $a_1x_1 + a_2x_2 + \dots + a_px_p = \underline{0}$ 가 성립하려면 모든 $a_i = 0$ 일 때만 만족한다면 열벡터 $(x_1, x_2, \dots, x_p)$ 는 선형 독립(linearly independent) 벡터라 함
- 0이 아닌 a_i 에 대해서 만족한다면 선형 종속(linearly dependent)인 벡터
- 상호 종속인 벡터는 하나의 벡터를 다른 벡터들의 선형 결합으로 표시할 수 있다는 것을 의미한다.

▪ 역행렬이 존재한다. ▪ full-rank이다. rank(A)=n ▪ A는 non-singular이다. ▪ $A \neq 0$ ▪ $Ax=b$ 의 해가 존재한다.	▪ 역행렬이 존재하지 않는다 ▪ full-rank아니다. rank(A)<n ▪ A는 singular이다. ▪ $A =0$ ▪ $Ax=b$ 의 해가 존재하지 않는다.
---	---

$B_{3 \times 2} = \begin{bmatrix} 1 & -1 \\ 2 & 0 \\ 1 & 2 \end{bmatrix} \Rightarrow rank(B) = 2,$

$M_{3 \times 3} = \begin{bmatrix} 1 & -1 & 0 \\ 2 & 0 & 2 \\ 1 & 2 & 3 \end{bmatrix} \Rightarrow rank(M) = 2$ 왜? 첫째 열과 둘째 열의 합은 3째열이다. 그러므로 3번째 열은 종속되어 있음 => 역행렬 M^{-1} 은 존재하지 않으며 $M\underline{x} = \underline{0}$ 을 만족하는 해는 무수히 많다.

행렬 활용

연립방정식 풀이

$$\begin{aligned} 2x - 2y - z &= 5 \\ x + y - 2z &= 1 \\ x - z &= 4 \end{aligned} \Rightarrow A\underline{x} = \underline{b}, A = \begin{bmatrix} 2 & -2 & -1 \\ 1 & 1 & -2 \\ 1 & 0 & -1 \end{bmatrix}, \underline{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \underline{b} = \begin{bmatrix} 5 \\ 1 \\ 4 \end{bmatrix} \Rightarrow (\text{해}) \Rightarrow \underline{x} = A^{-1} \underline{b}$$

일변량 확률변수 x 기대값 행렬 표현

확률변수 x의 n개 관측치 벡터를 $\underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ 이라 하자.

- $\sum x_i \Rightarrow$ (행렬 표현) $\sum x_i = (\underline{1}'\underline{x})$, where $\underline{1}$ 은 차수가 n인 1벡터, 모든 원소가 1

- $\sum x_i^2 \Rightarrow \sum x_i^2 = \underline{x}'\underline{x} \quad / \quad \text{평균 } \bar{x} = \sum x_i/n = (\underline{1}'\underline{x})/n$

확률변수 벡터 $\underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_p \end{bmatrix}$

확률변수 x_i 의 평균과 분산을 각각 μ_i, σ_{ii}^2 라 하고 (x_i, x_j) 의 공분산을 σ_{ij}^2 라 하면

$$(1) E(\underline{x}) = \underline{\mu} = \begin{bmatrix} \mu_1 \\ \dots \\ \mu_p \end{bmatrix} \quad (2) V(\underline{x}) = \Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \dots & \sigma_{1p}^2 \\ \sigma_{11}^2 & \sigma_{22}^2 & \dots & \sigma_{2p}^2 \\ \dots & \dots & \dots & \dots \\ \sigma_{p1}^2 & \sigma_{p2}^2 & \dots & \sigma_{pp}^2 \end{bmatrix} \text{이다.}$$

(3) $E(\underline{a}'\underline{x}) = \underline{a}'\underline{\mu}$: 결과는 스칼라 (4) $V(\underline{a}'\underline{x}) = \underline{a}'\Sigma\underline{a}$

선형회귀모형(예측변수 p개)

- $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi} + e_i, \text{ for } i = 1, 2, \dots, n, e_i \sim iidN(0, \sigma^2)$
- 행렬표현 $\underline{y}_{n \times 1} = X_{n \times (p+1)} \underline{b}_{(p+1) \times 1} + \underline{e}_{n \times 1}, \underline{e} \sim N(\underline{0}, \sigma^2 I)$
- 평균 : $\bar{y} = \sum y_i/n = [11\dots 1]\underline{y} = \underline{1}'\underline{y}/n$
- 분산 : $s(y) = \sum (y_i - \bar{y})^2/(n-1) = (\underline{y} - \underline{\mu})'(\underline{y} - \underline{\mu})/(n-1)$

행렬의 미분

상수벡터 $\underline{a}' = [a_1 \ a_2 \ \dots \ a_p]$, 정방행렬 A 에 대하여

(1) $\frac{\partial}{\partial \underline{x}}(\underline{a}'\underline{x}) = \underline{a}$ (2) $\frac{\partial}{\partial \underline{x}}(\underline{x}'\underline{a}) = \underline{a}$ (3) $\frac{\partial}{\partial \underline{x}}(\underline{x}'A\underline{x}) = A\underline{x} + A'\underline{x} \Rightarrow (A \text{ 대칭행렬}) \frac{\partial}{\partial \underline{x}}(\underline{x}'A\underline{x}) = 2A\underline{x}$

회귀모형 $\underline{y}_{n \times 1} = X_{n \times (p+1)} \underline{b}_{(p+1) \times 1} + \underline{e}_{n \times 1}$

OLS 추정 : $\min_{\alpha, \beta_1, \dots, \beta_p} \sum e_i^2 = \underline{e}'\underline{e} = (\underline{y} - X\underline{b})'(\underline{y} - X\underline{b})$

고유치 eigen value 고유벡터

고유치 구하기

- 차수가 n 인 정방행렬 $A_{n \times n}$ 의 $|A - \lambda I|$ 행렬식을 계산한다. 이 방정식은 차수가 n 인 λ 의 다항방정식이다. $f(\lambda) = a_1\lambda^n + a_2\lambda^{n-1} + \dots$
- $f(\lambda) = a_1\lambda^n + a_2\lambda^{n-1} + \dots = 0$ 특성방정식(characteristic equation)를 만족하는 $\lambda_1, \lambda_2, \dots, \lambda_n$ (n 개 존재) 고유치(eigen value, characteristic value, latent value)라 한다.
- 이렇게 구한 λ_i 는 $A - \lambda I$ 행렬을 singular로 만든다

고유벡터 구하기

- 각 λ_i 에 대하여 $(A - \lambda I)\underline{x} = \underline{0}$ 을 구하면 이를 고유벡터라 한다.
- [단점] $A - \lambda I$ 행렬을 singular이므로 고유벡터 $\underline{x}$ 는 무수히 많이 존재한다.
- [장점] 이 성질 때문에 요인 분석에서 요인 회전이 가능하다.
- 활용 시 $\underline{x}'\underline{x} = 1$ 을 만족하는 정규(Norm) 고유벡터를 사용한다. $\underline{e}_i$ - 정규 고유벡터 기호

성질

- 고유치의 합은 행렬 A 의 대각원소의 합이다. $trace(A) = \sum \lambda_i$
- 고유벡터는 서로 독립이다. => 좌표를 독립(상호 독립, 직교) 공간에 표현한다.

