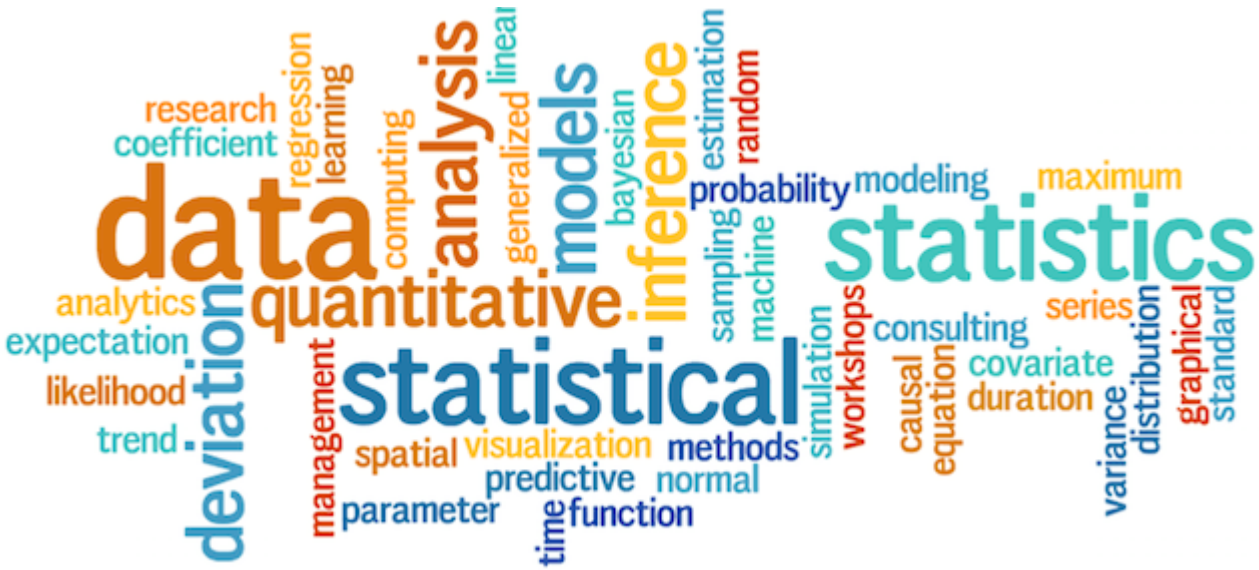


통계기반 데이터 분석 (Ch 4)

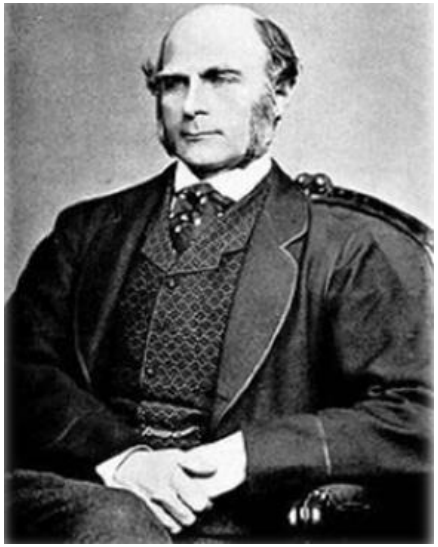
서강대학교 정보통신대학원 데이터사이언스 & A.I
정화민 교수

Regression

Regression 의 역사

- 회귀(regress 리그레스)의 원래 의미는 옛날 상태로 돌아가는 것을 의미함.
- 영국의 유전학자 프랜시스 골턴은 부모의 키와 아이들의 키 사이의 연관 관계를 연구하면서 부모와 자녀의 키 사이에는 선형적인 관계가 있고 키가 커지거나 작아지는 것보다는 전체 키 평균으로 돌아가려는 경향이 있다는 가설을 세웠으며 이를 분석하는 방법을 "회귀분석"이라고 하였다
- 이러한 경험적 연구 이후, 칼 피어슨은 아버지와 아들의 키를 조사한 결과를 바탕으로 함수 관계를 도출하여 회귀분석 이론을 수학적으로 정립하였다.

Source: www.wikipedia.org



Francis Galton(1886)

Karl Pearson(1903)

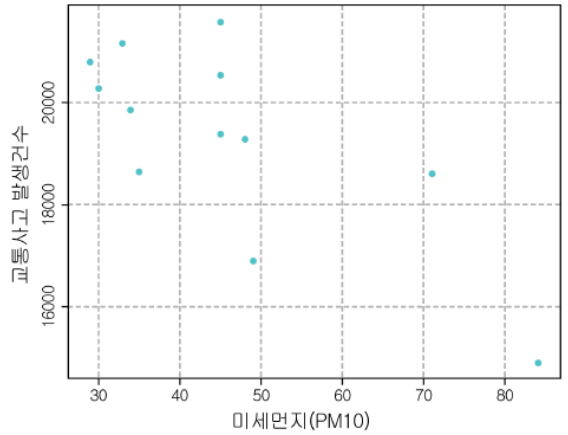
Regression

- 독립변수와 종속변수를 구별하고 인과관계를 파악
- 통계적 모형 구축의 예로 단순선형회귀분석의 과정을 이해
- 회귀분석의 가정을 만족하는지 확인
- 인과관계 :

  원인과 결과 관계를 뜻하는 인과관계는 주의 깊은 자료의 관찰을 통해 얻을 수 있는 관계

  데이터에 대한 김은 통찰과 경험이 있어야 되며 통찰이 없다면 잘못된 인과관계 도출될 수 있다.

예) 아래 도표는 미세먼지 농도가 짙어질수록 교통사고 발생이 줄어드는 경향을 보이고 있다. 그러면 이를 통헤 미세먼지가 증가하면 교통사고 건수가 줄어든다고 할 수 있는가 ?



미세먼지와 교통사고 발생 도표

Regression

Regression 의 변수

- 사회현상에 대한 관찰 연구가 필요한 경우가 많음
- 현상을 이해하기 위하여 사전지식과 사회에 대한 깊은 통찰력을 가져야 한다. (변수의 연관성 고려, 원인과 결과에 대한 고민, 제 3의 요인 등)

사례 : 1) 아이스크림 판매량이 증가할 수록 익사사고 발생이 증가하였다. 익사사고 발생을 억제하기 위해 아이스크림 판매를 금지해야 하는 것이 올바른 결정인지 ?(제 3의 요인 : 계절)

2) 불을 켜고 자는 어린이의 경우, 나이가 들어 근시가 될 경우가 많다. 근시 예방을 위하여 어릴 때 부터 잠을 잘 때 불을 꺼고 자야하는 것이 올바른 결정인지 ? (제 3의 요인 : 유전적 영향)

3) 국가 부채가 GDP의 90% 이상이 될 경우 국가의 성장률이 느려진다. 높은 국가 부채는 국가의 성장을 느리게 한다가 맞는지 ? (뒤바뀐 인과관계)

4) 사과의 수입이 증가할 수 록 이혼률이 증가한다. 이혼률을 낮추기 위해 사과 수입을 금지해야 올바른 결정인지 ? (인과관계를 알수 없는 경우)

- 종속변수: 변수에 의해 영향을 받아 그 값이 결정되는 변수
- 독립변수: 영향을 미치는 변수를 독립변수

Regression

Simple Regression Analysis

- 독립변수 하나, 종속변수 하나,
- 원인과 결과를 통계적으로 검정
- 해석방법 : 통계적으로 유의할 시, 독립변수(원인)은 종속변수(결과)에 통계적으로 유의한 영향을 준다.

  (+)면 긍정적 , (-)면 부정적 영향
- 단순회귀모형의 추정 (찾아야할 내용: Y 절편, 기울기 b(힘의크기= 영향력), 결정계수 R^2 (정확도))
- 원리 파악 : 엑셀 이용 (추정치 , 잔차)

	A	B	C	D	E
1	번호	노동력(X)	생산량(Y)	X^2	XY
2	1	267	428	71,289	114,276
3	2	263	430	69,169	113,090
4	3	238	417	56,644	99,246
5	4	219	384	47,961	84,096
6	5	274	432	75,076	118,368
7	6	257	425	66,049	109,225
8	7	321	474	103,041	152,154
9	8	305	462	93,025	140,910
10	9	285	449	81,225	127,965
11	10	247	405	61,009	100,035
12	합계	2,676	4,306	724,488	1,159,365
13		A	B	C	D
14					
15					
16	N =	10			
17					
18	a =	204.8125			
19					
20	b =	0.8438			
21					

$$Y = a + bX$$

$$a = \frac{BC - AD}{nC - A^2}$$

$$b = \frac{nD - AB}{nC - A^2}$$

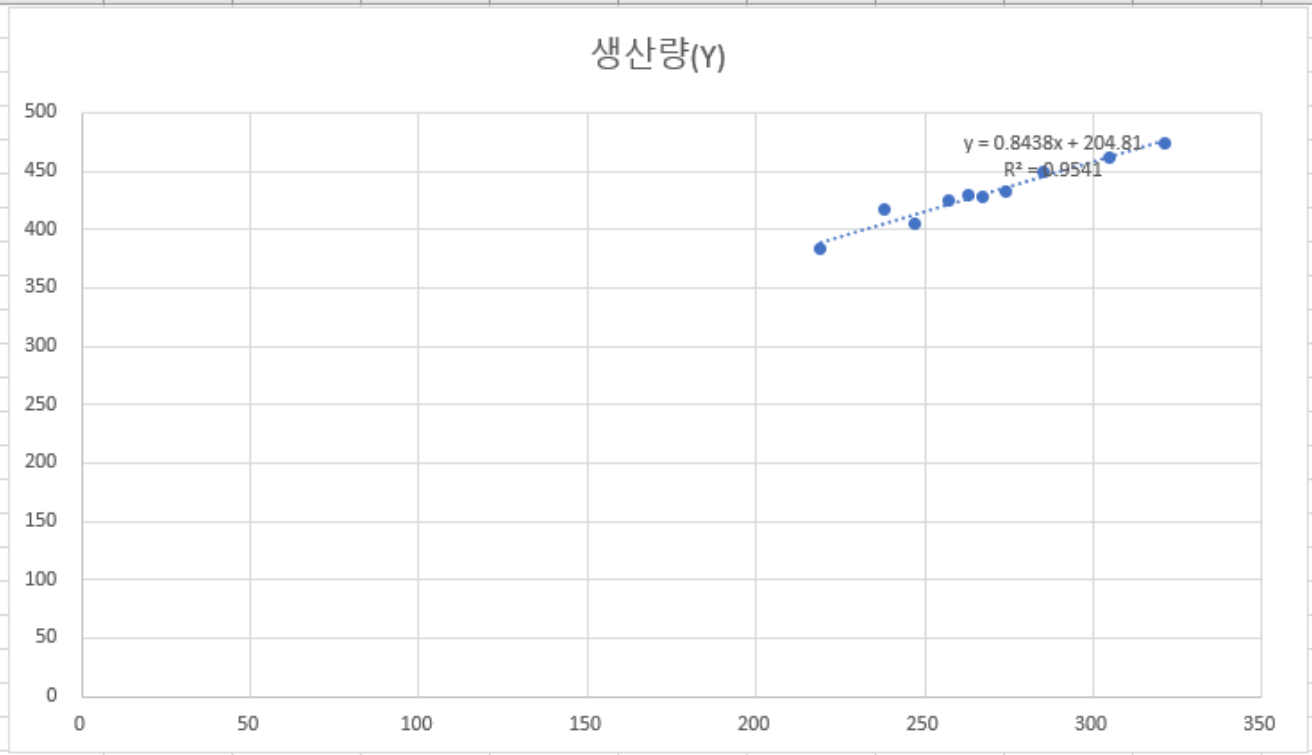
	A	B	C	D	E	F	G	H
1	번호	노동력(X)	생산량(Y)	X^2	XY	추정치	잔차	
2	1	267	428	71,289	114,276	430	-2	
3	2	263	430	69,169	113,090	427	3	
4	3	238	417	56,644	99,246	406	11	
5	4	219	384	47,961	84,096	390	-6	
6	5	274	432	75,076	118,368	436	-4	
7	6	257	425	66,049	109,225	422	3	
8	7	321	474	103,041	152,154	476	-2	
9	8	305	462	93,025	140,910	462	0	
10	9	285	449	81,225	127,965	445	4	
11	10	247	405	61,009	100,035	413	-8	
12	합계	2,676	4,306	724,488	1,159,365			
13								
14								
15				Y 분산	626.0400	←	=VARP(C2:C11)	
16	N =	10						
17				추정치분산	597.3244	←	=VARP(F2:F11)	
18	a =	204.8125						
19				잔차분산	28.7156	←	=VARP(G2:G11)	
20	b =	0.8438						
21				R^2	0.9541	←	=1 -(E19/E15)	

Regression

Simple Regression Analysis

- 산점도를 이용한 회귀계수 구하기
- 데이터에 전체(독립변수, 종속변수 블록) -> 삽입 -> 차트 -> 분산형 -> 추세선 –선형 선택 (수식차트에 표시, R 제곱값 차트에 표시)

	A	B
1	노동력(X)	생산량(Y)
2	267	428
3	263	430
4	238	417
5	219	384
6	274	432
7	257	425
8	321	474
9	305	462
10	285	449
11	247	405



Regression

▪ Simple Regression Analysis

- 엑셀 추가기능 확장(분석도구 √)
- 데이터 -> 데이터분석 -> 회귀분석
- F 검정 (F-test) : 회귀식 전체에 대한 유의성 검정($p<0.05$ 일 때 유의함)

-> 모든 회귀계수가 "0"이라는 귀무가설의 기각여부를 검정하는 것으로 귀무가설이 기각되고 대립가설이 채택된다는 의미

- T검정 (T-test) : 각 독립변수가 개별적으로 유의한 지를 보고자 하는 것임 ($p<0.05$)

-> t 검정을 하는 경우 95% 신뢰수준을 가정 유의수준을 5% , 이와 같은 가설검정하는 방법을 양측검정 (two - tailed test)

	A	B	C	D	E	F	G	H	I
1	요약 출력								
2									
3	회귀분석 통계량								
4	다중 상관계수	0.976796461							
5	결정계수	0.954131325							
6	조정된 결정계수	0.948397741							
7	표준 오차	5.991204491							
8	관측수	10							
9									
10	분산 분석								
11		자유도	제곱합	제곱 평균	F 비	유의한 F			
12	회귀	1	5973.24375	5973.24375	166.4109696	1.23325E-06			
13	잔차	8	287.15625	35.89453125					
14	계	9	6260.4						
15									
16		계수	표준 오차	t 통계량	P-값	하위 95%	상위 95%	하위 95.0%	상위 95.0%
17	Y 절편	204.8125	17.60509644	11.63370509	2.71421E-06	164.2150748	245.4099252	164.2150748	245.4099252
18	노동력(X)	0.84375	0.065406786	12.90003758	1.23325E-06	0.692921681	0.994578319	0.692921681	0.994578319
19									

Multiple Regression

단순회귀 분석

종업원	연봉(Y)	연수(X)
1	325	2
2	275	4
3	350	4
4	400	6
5	325	7
6	425	8
7	375	10
8	475	10
9	400	12
10	575	12
11	425	14
12	450	16
13	700	17
14	525	18
15	600	20
16	750	20
17	650	22
18	775	23
19	675	24
20	825	26

공식 $Y = a + bX$

회귀공식과 결정계수 비교하기

수업시간에 실습

(학력 1: 대졸, 학력 0 고졸)

1 . 단순회귀와 다중회귀분석의 결정계수 비교

2. 회귀공식 완성하기

다중회귀분석

종업원	연봉(Y)	연수(X)	학력(W)
1	325	2	1
2	275	4	0
3	350	4	1
4	400	6	1
5	325	7	0
6	425	8	1
7	375	10	0
8	475	10	1
9	400	12	0
10	575	12	1
11	425	14	0
12	450	16	0
13	700	17	1
14	525	18	0
15	600	20	0
16	750	20	1
17	650	22	0
18	775	23	1
19	675	24	0
20	825	26	1

공식 $Y = a + bX + cW$

Mutiple Regression

다중회귀로 분석한 결정계수(0.97) 가 단순회귀분석의 결정계수(0.81)보다 우수함

요약 출력

회귀분석 통계량	
다중 상관계	0.903091294
결정계수	0.815573885
조정된 결정	0.80532799
표준 오차	73.89899983
관측수	20

분산 분석

	자유도	제곱합	제곱 평균	F 비	유의한 F
회귀	1	434700.88	434700.88	79.60006073	5.01364E-08
잔차	18	98299.119	5461.0622		
계	19	533000			

	계수	표준 오차	t 통계량	P-값	하위 95%	상위 95%	하위 95.0%	상위 95.0%
Y 절편	231.3873257	35.82676	6.4585055	4.46785E-06	156.1180962	306.65656	156.1181	306.65656
연수(X)	20.62637632	2.3118851	8.9218866	5.01364E-08	15.76928605	25.483467	15.769286	25.483467

요약 출력

회귀분석 통계량	
다중 상관계	0.987399067
결정계수	0.974956918
조정된 결정	0.972010673
표준 오차	28.02096121
관측수	20

분산 분석

	자유도	제곱합	제곱 평균	F 비	유의한 F
회귀	2	519652.04	259826.02	330.915097	2.4482E-14
잔차	17	13347.963	785.17427		
계	19	533000			

	계수	표준 오차	t 통계량	P-값	하위 95%	상위 95%	하위 95.0%	상위 95.0%
Y 절편	148.8173757	15.734048	9.4583017	3.48024E-08	115.6214355	182.01332	115.62144	182.01332
연수(X)	21.84915811	0.8844658	24.703224	9.25206E-15	19.9830983	23.715218	19.983098	23.715218
학력(W)	131.5134004	12.643531	10.401635	8.66663E-09	104.8378811	158.18892	104.83788	158.18892

Multiple Regression

함수는 특정한 작업을 수행하기 위해 일련의 구문들을 체계적으로 묶은 것, R은 수많은 내장 함수를 가지고 있음.

사용 Data : R 내장 데이터 mtcars:
1974년 미국 Motor trend US magazine에 나오는 data이다.
32개의 차량에 대해서 각 자료들을 기재되어 있음

mpg - Miles/gallon (연비, 1갤런당 몇 마일을 가는가)
cyl - Number of cylinders (차량 엔진의 실린더의 개수, 펌프같이 움직이는 것)
disp - Displacement (배기량)
hp - Gross horsepower (마력)
drat - Rear axle ratio (후방 축 비율
wt - Weight (1000lbs) 파운드 기준 차량무게
qseq - 1/4 mile time 1/4 마일 간 시간?
am - Transmission(0 = automatic, 1 = manual) 변속기가 자동이냐 아니냐
gear - Number of forward gears 전진기어의 수? (1,2,3 **)**
carb - Number of carburetors 카뷰레이터 수 (기화기수**)**

Multiple Regression

연구문제 : 다중회귀분석을 이용하여 주행연비 계산하기 ?
회귀식 만들기 ? y = a + b1x1 + b2x2 +...bnxn
종속변수: 연비(mpg), 독립변수: 배기량(disp), 마력(hp), 무게(wt)

```
input <- mtcars[,c("mpg","disp","hp","wt")]
print(head(input))
```

```
                   mpg disp  hp    wt
Mazda RX4         21.0  160 110 2.620
Mazda RX4 Wag     21.0  160 110 2.875
Datsun 710        22.8  108  93 2.320
Hornet 4 Drive    21.4  258 110 3.215
Hornet Sportabout 18.7  360 175 3.440
Valiant           18.1  225 105 3.460
```

```
input <- mtcars[,c("mpg","disp","hp","wt")]
 model <- lm(mpg~disp+hp+wt, data = input)
print(model)
```

```
Call:
lm(formula = mpg ~ disp + hp + wt, data = input)

Coefficients:
(Intercept)         disp           hp           wt
  37.105505    -0.000937    -0.031157    -3.800891
```

y = a + b1x1 + b2x2 +...bnxn

아래 공식을 완성하세요

Y = 37.10 +(-0.000937*200)+
(-0.031157*120)+(-3.800891*2.91)