1

개념

1. 정의 Definition

2개 이상 (양적) 변수 데이터

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

개체(subject)

변수(variable)

- 인과 관계(casual relationship)를 규명, 분석하거나 (회귀 분석: regression, 다변량 분산 분석: Multivariate ANOVA) => 예측모형
- 변수들의 상관관계(correlation)를 이용하여 변수의 차원을 축약(reduction)하거나
- 개체 간의 유사성을 측정하여 개체 분류(classification)에 관련된 분석 방법

일반적으로 (협의의) 다변량 분석은 1) 차원 축소, 2) 개체 분류를 의미함.

2. 확률변수 randpm variable)

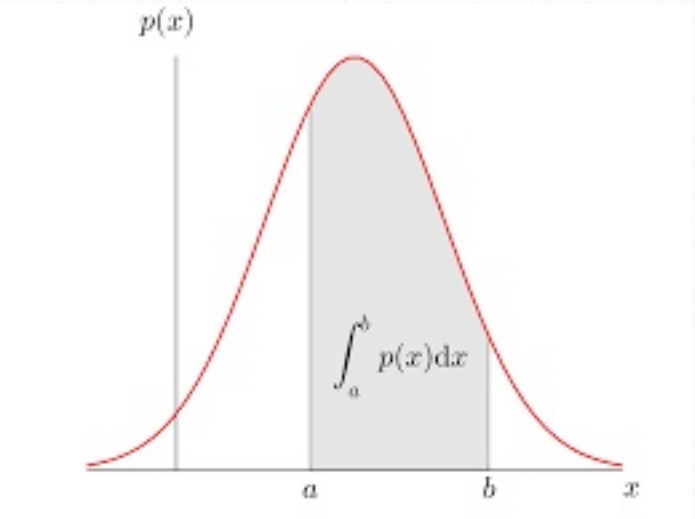
1) 정의

확률실험의 결과(원소 w)와 실수 값을 대응한 함수 $X(w) = x$

- 데이터에서 확률변수는 개체의 관심 특성의 총칭이며 변수의 관측값이 데이터
- 실험결과가 측정형인 경우 결과가 확률변수임. (예) 주가 관측실험에서 결과는 주가 실수이므로 결과 자체가 확률변수, 동전을 던지는 실험에서 앞면=1, 뒷면=0으로 변환해 주는 함수를 확률변수라 함

2) 확률분포함수

확률변수가 가지는 값과 그에 대응하는 확률을 표, 수식, 그래프로 나타낸 함수 ($f(x)$)



- 확률분포함수는 데이터에 대한 모든 정보를 갖는다.
- 데이터 관측값들의 평균, 최대값, 최소값, 일정 구간(범위)에 속할 확률
- 표본 데이터의 히스토그램이 확률분포함수이다.
- 데이터의 히스토그램으로 확률변수함수 식을 도출하는 것은 불가능하므로 일반적으로 가정하게 됨 - 통계분석에서 가정되는 분포는 정규분포, 적어도 좌우대칭 분포임

3) 변수 종류 : 분석적 측면

1) Metric (측정형 변수, measurable) : 실험 개체의 측정 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. 예) 교통량

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

- 명목형 (nominal): 분류만 ➜ 성별, 결혼여부, 소득(단위: 만원)
- 순서형 (ordinal): 순서를 가진 분류 ➜ 성적(A, B, ..) 소득수준(상, 중, 하), 5점 척도

대부분의 다변량 분석은 변수들의 분포가 다변량 정규 분포(multivariate normal distribution)를 가정하기 때문에 고전적 다변량 분석은 metric 변수들에 분석이나 non-metric 변수 중 순서형 변수에 대한 분석 방법도 제안되고 있다.

4) 변수종류 : 시간에 의해

자료가 시간적 순서를 가지면 이를 시계열 자료(time series)라 하고 그렇지 않은 경우를 횡단면 자료(Cross-section: 일정 시간에 한꺼번에 조사)라 한다.

경제 지표(환율, 수출량)나 기업의 연차별 자료(연도별 매출액)가 시계열 자료에 해당된다. 시계열 다변량 자료에 대한 분석 방법으로는 시계열 자료 분석, 계량 경제(econometrics) 방법을 이용하면 된다.

5) 변수종류 : 인과 관계

인과 관계(casual relationship)에서 원인이 되는(영향을 주는) 변수를 설명 변수(exploratory var.) 혹은 독립변수(independent var.)라 하고 결과나 영향을 받는 변수를 종속 변수(dependent var.) 혹은 반응 변수

(response var.)라 한다. 종속 변수는 Y, 설명 변수는 X로 표시한다. 분산 분석에서는 설명 변수를 처리 효과, 요인으로 불리어진다.

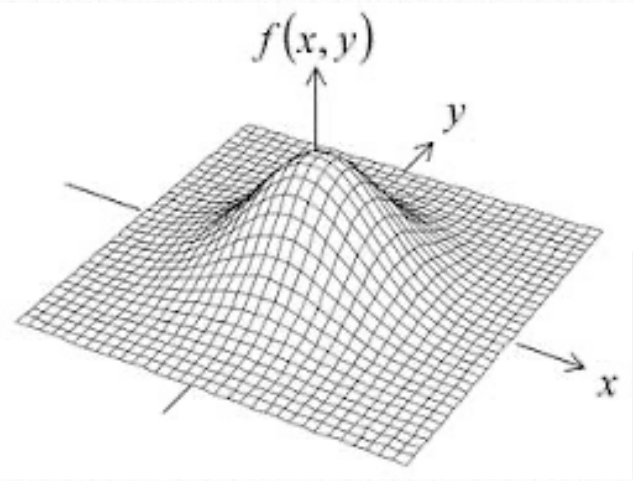
인과 관계는 이론적, 경험적 타당성에 근거하여 연구 목적에 설정되는 것이지 자료 분석 후 인과 관계가 설정되는 것은 아니다.

3. 다변량분포

확률변수가 p(≥ 2)개인 결합확률분포함수, 다변량분석에서는 변수들이 좌우 대칭인 분포를 따르는 것이 적합하다. 그렇지 않으면 적절한 정규변환이 필요함 정규변환강의가기

1) 이변량 정규분포

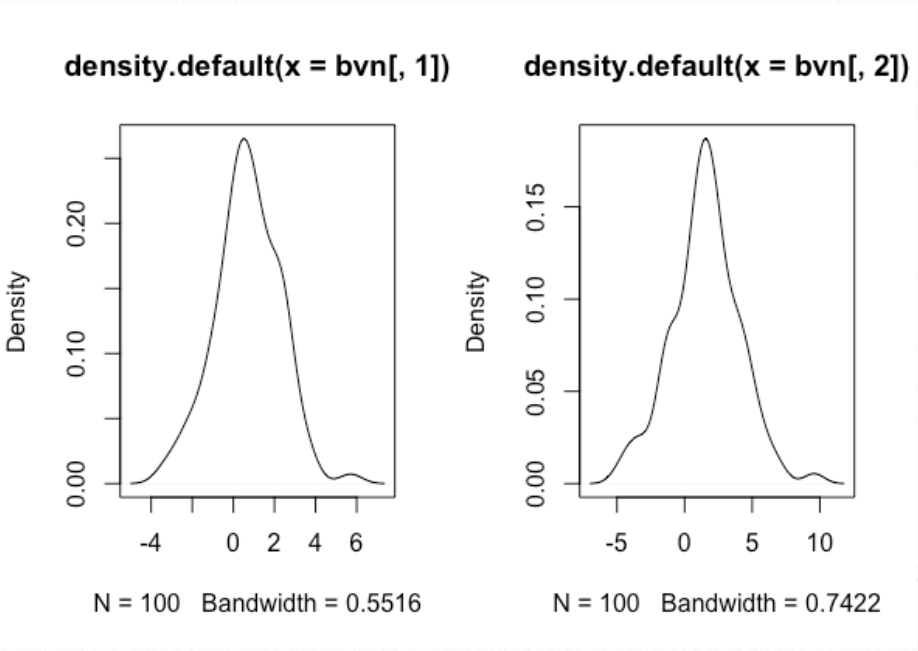
확률변수 (x, y)가 이변량 정규분포를 따른다고 하면 결합확률분포함수 그래프는 다음과 같다.



$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \cdot \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right\}$$

(정리) 만약 $Z_1, Z_2 \sim iid\ n(0,1)$이면, $X = \sigma_X Z_1 + \mu_X$, $Y = \sigma_Y[\rho Z_1 + \sqrt{1-\rho^2} Z_2] + \mu_Y$ ~ BVN

```
library(MASS) #mvnorm function package
Sigma <- matrix(c(2,3,3,5),2,2) #covariance matrix
bvn<-as.matrix(mvrnorm(n = 100, c(1, 2), Sigma))
par(mfrow=c(1,2))  #plot 2 in 1 row
plot(density(bvn[,1])) #marginal of X
plot(density(bvn[,2])) #marginal of Y
```



이변량 정규분포의 주변확률분포는 정규분포를 따른다.

2) 다변량 정규분포

차수 p인 확률변수 벡터 $\underline{x}_p = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$

$\underline{x}_p \sim N_p(\underline{\mu}_p, \Sigma_{p\times p})$, 평균이 $\underline{\mu}$, 공분산행렬이 Σ인 다변량 정규분포의 확률분포함수는 다음과 같다.

$$f_{\underline{x}}(\underline{x};\underline{\mu},\Sigma) = \frac{1}{(2\pi)^{p/2}\,|\Sigma|^{1/2}} \exp[-1/2(\underline{x}-\underline{\mu})'\Sigma^{-1}(\underline{x}-\underline{\mu})]$$

4.고유값 eigen value [강의가기]

- 데이터 변수의 변동의 크기, 데이터의 공분산으로부터 구한 고유값은 원 변수들의 총변동의 크기를 반영함
- 변수의 변동(분산)은 관심 개체를 구별하는 능력 - 변수의 변동이 크다는 것은 개채를 구별할 수 있는 능력이 크다는 것임
- 예를 들어 H대학 통계학과 학생들의 고등학교 수학 성적과 영어성적을 조사한 자료이다.
- 수학평균=70점, 분산=20점 - 영어평균=80, 분산=5점
- 어느 과목을 이용하여 학생들의 능력을 구별하는 것이 좋을까? 당연히 변동(분산)이 큰 수학성적으로 학생들을 구별해야 용이하게 구별할 수 있음

5. 다변량 분석 요약

	주성분	요인	판별	군집	정준상관
변수 관계 탐색	S	D	N	N	S
자료 탐색	D	S	N	S	N
새 변수 만들기	Yes	Yes	No	No	Yes
개체 분류	No	No	Yes	Yes	No
변수 그룹	P	P	N	N	D
차원 줄이기	D	P	N	N	N

Sometimes, Definitely, Never, Possible, Rarely

I다변량 분석 개념도I

미국 59개 도시들의 기후지표, 사회지표, 환경지표에 대한 데이터

<- 기후지표-> <- 사회지표-> <- 환경지표->

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	도시이름	1월기온	7월기온	상대습도	강수량	사망률지	교육기간	인구밀도	비백인비	사무직종	총인구	가구당인	가계소득	HCPot	NOxPot	SO2Pot	NOx	남부여부
2	Akron, OH	27	71	59	36	921.87	11.4	3243	8.8	42.6	660328	3.34	29560	21	15	59	15	NO
3	Albany-S	23	72	57	35	997.87	11	4281	3.5	50.7	835880	3.14	31458	8	10	39	10	NO
4	Allentown	29	74	54	44	962.35	9.8	4260	0.8	39.4	635481	3.21	31856	6	6	33	6	NO
5	Atlanta, G	45	79	56	47	982.29	11.1	3125	27.1	50.2	2E+06	3.41	32452	18	8	24	8	YES
6	Baltimore,	35	77	55	43	1071.3	9.6	6441	24.4	43.7	2E+06	3.44	32368	43	38	206	38	YES
7	Birmingha	45	80	54	53	1030.4	10.2	3325	38.5	43.1	883946	3.45	27835	30	32	72	32	YES

분석목적 (1) 59개 도시를 측정된 지표들을 이용하여 특성에 따른 분류
(2) 측정된 지표들을 이용하여 “남부여부”를 판별하는 규칙(패턴) 탐색

1) 분석 (1) & (2)는 개체(도시)에 대한 분석임 -> 이를 개체 관련 기법 subject directed tech.

2) 특성 지표 3개분야, 총 15개 변수를 이용하여 도시의 특성을 파악하기는 변수의 수가 너무 많다. - 변수가 1개이면 기초통계량, 2개면 산점도, 3개면 버블산점도로 특성 파악 가능, 그러나 4개 이상이면 변수의 차원을 줄여야 도시 특성 파악이 가능하다.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{21} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{n1} \end{bmatrix}$$

개체분류

- 군집분석 clustering analysis : 유사한 개체들을 동일 군집 분류
- 판별분석 discriminant analysis : 새로운 개체를 가장 적절한 특정 집단으로 분류

차원축약

- 주성분분석 principle component analysis : 변수들의 상관관계를 활용하여 변수 개수 축약 – 새로운 변수, 주성분 변수
- 요인분석 factor analysis : 유사한 변수들을 군집으로 분류

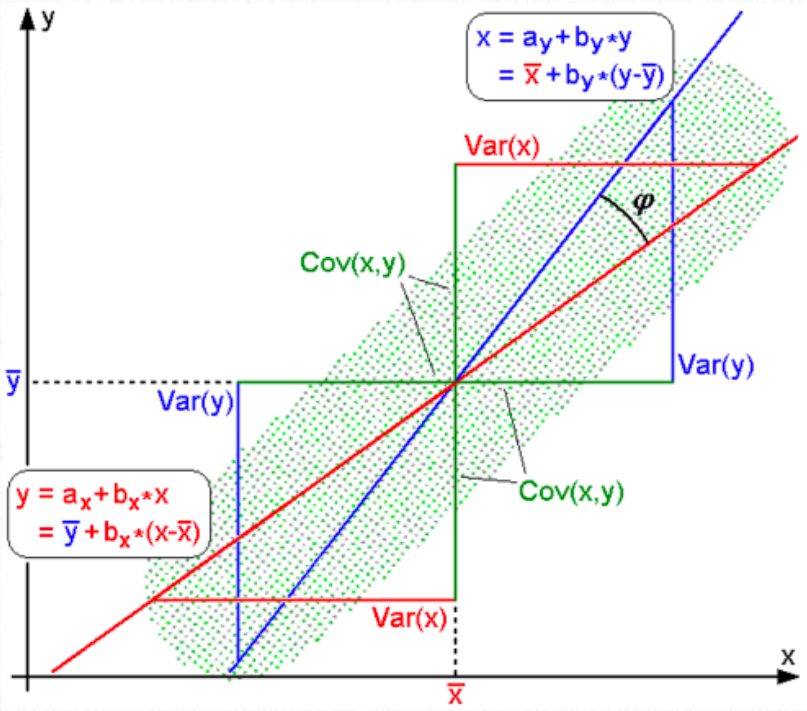
2

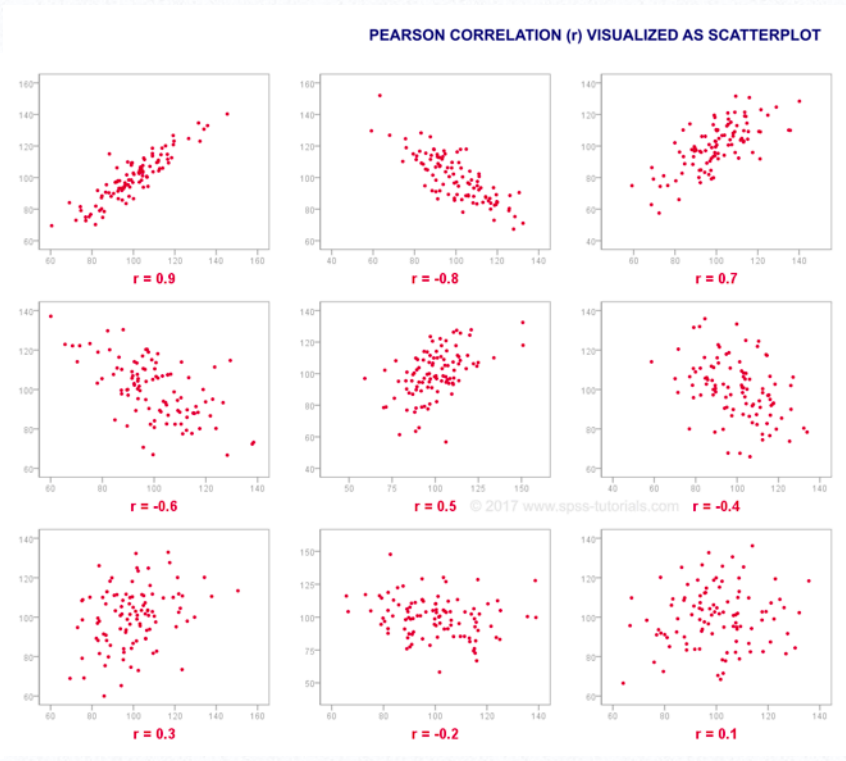
상관분석

1. 개념

두 측정형 변수 간의 직선적 관계 정도를 파악한다.

[기하학적 의미] 출처 - 위키피디아





2. 상관계수 공식

데이터 $(x_i, y_i), i = 1,2,...,n$

1) pearson 상관계수

공식 $r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$

검정

- 귀무가설 : $\rho_{xy} = 0$ 두 확률변수는 서로 독립이다.
- 검정통계량 : $TS = r\sqrt{\frac{n-2}{1-r^2}} \sim t(n-2)$ - 두 변수가 이변량 정규분포를 따른다면,
- $TS = \frac{t}{\sqrt{n-2+t^2}} \sim N(0,1)$ - 두 변수가 이변량 정규분포를 따르지 않고 n이 충분히 크다면,

신뢰구간 구하기

- $F(r) = \frac{1}{2} ln \frac{1+r}{1-r}$, $F(\rho) = \frac{1}{2} ln \frac{1+\rho}{1-\rho}$
- 검정통계량 $[F(r) - F(\rho)]\sqrt{n-3} \sim N(0,1)$
- 귀무가설 $H_0 : \rho = \rho_0 \neq 0$ 검정을 위하여, 검정통계량 $[F(r) - F(\rho_0)]\sqrt{n-3} \sim N(0,1)$

2) spearman 순위 rank 상관계수

- R_{x_i} - i번째 X 관측치의 순위
- R_{y_i} - i번째 Y 관측치의 순위
- $d_i = R_{x_i} - R_{y_i}$: i-번째 (X, Y) 관측치 순위 차이
- 검정통계량 $TS = 1 - \frac{6\sum_i^n d_i^2}{n(n^2-1)}$
- 유의성 검정은 pearson 상관계수 유의성 검정과 동일한 방법, 혹은 permutation test 사용

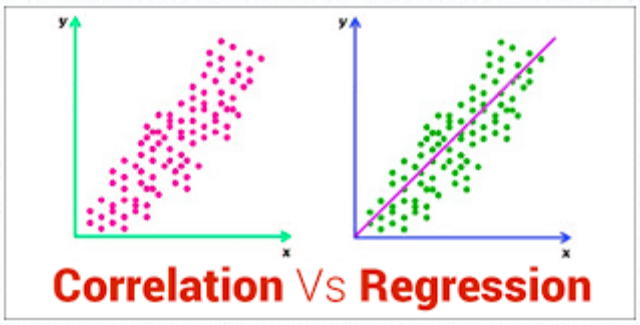
3) Kendall's Tau 순위 상관계수

서로 다른 두 쌍의 관측치 (x_i, y_i), (x_j, y_j)에 대하여

- $x_i > x_j \ and \ y_i > y_j$ or $x_i < x_j \ and \ y_i < y_j$ - concordant pair
- $x_i > x_j \ and \ y_i < y_j$ or $x_i < x_j \ and \ y_i > y_j$ - disconcordant pair
- 그외 경우는 condordant, disconcordant 아님
- 검정통계량 $\tau = \frac{\#\ of\ concordantpairs - \#\ of\ disconcordantpairs}{nC_2} \sim N(0, \frac{2(2n+5)}{9n(n-1)})$

3. 회귀분석과 관계

1) 단순회귀모형 $Y_i = a + bX_i + e_i$



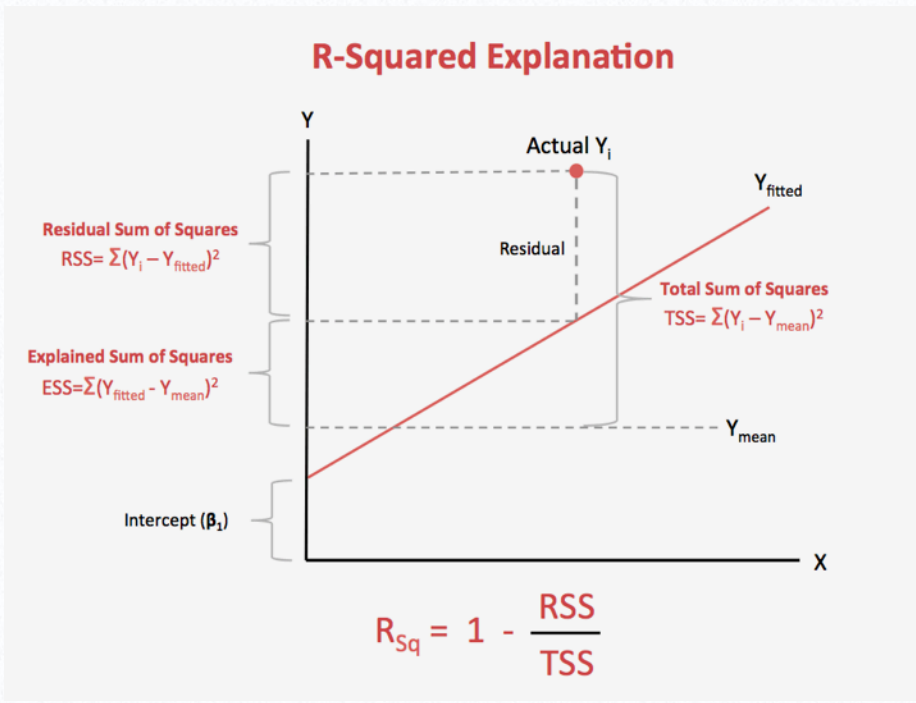
OLS 추정치 : $\hat{b} = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

- $H_0 : b = b_0$: 검정통계량 $TS = \dfrac{\hat{b} - b_0}{s(\hat{b})} \sim t(n-2),\ s(\hat{b}) = \sqrt{\dfrac{MSE}{\sum (x_i - \bar{x})^2}},\ MSE = \dfrac{\sum (y_i - \hat{\alpha} - \hat{\beta} x_i)^2}{(n-2)}$

관계 : $\hat{b} = r\sqrt{\dfrac{\sum (y_i - \bar{y})^2}{\sum (x_i - \bar{x})^2}}$ - 정비례관계

2) 중회귀모형 $Y_i = a + \sum_{j}^{p} b_j X_i j + e_i$

- 총변동 $TSS = \sum (y_i - \bar{y})^2$
- 오차변동 $RSS = \sum (y_i - \hat{y}_i)^2$
- 모형변동 $ESS = TSS - RSS = \sum (\hat{y}_i - \bar{y})^2$
- 결정계수 $R^2_{y.x_1 x_2 \ldots x_p} = \dfrac{ESS}{TSS} = 1 - \dfrac{RSS}{TSS}$
- 단순회귀모형 $\sqrt{R^2} = \pm r$



4. 상관계수와 산점도

다변량에서의 활용

similarity of variable : 상관계수가 높은 변수들은 유사하다는 의미임

- 두 변수가 유사하다는 것은 관심 개체에 대한 정보가 겹친다는 의미임
- (몸무게와 키) 두 변수의 상관계수가 매우 크므로 두 변수 중 하나의 관측값만 알아도 그 사람의 다른 관측값을 쉽게 예측할 수 있다. 사람들의 신체적 정보를 얻기 위하여 두 변수 모두 필요하는 것은 아니므로 변수의 차원을 줄일 수 있음
- (수학, 영어) 성적의 상관관계는 다소 낮음, 그러므로 수학성적만으로 영어성적을 예상하기는 쉽지 않으므로 학생(개체)의 능력을 알려면 두 변수 모두 조사할 필요가 있음

5. 예제 데이터

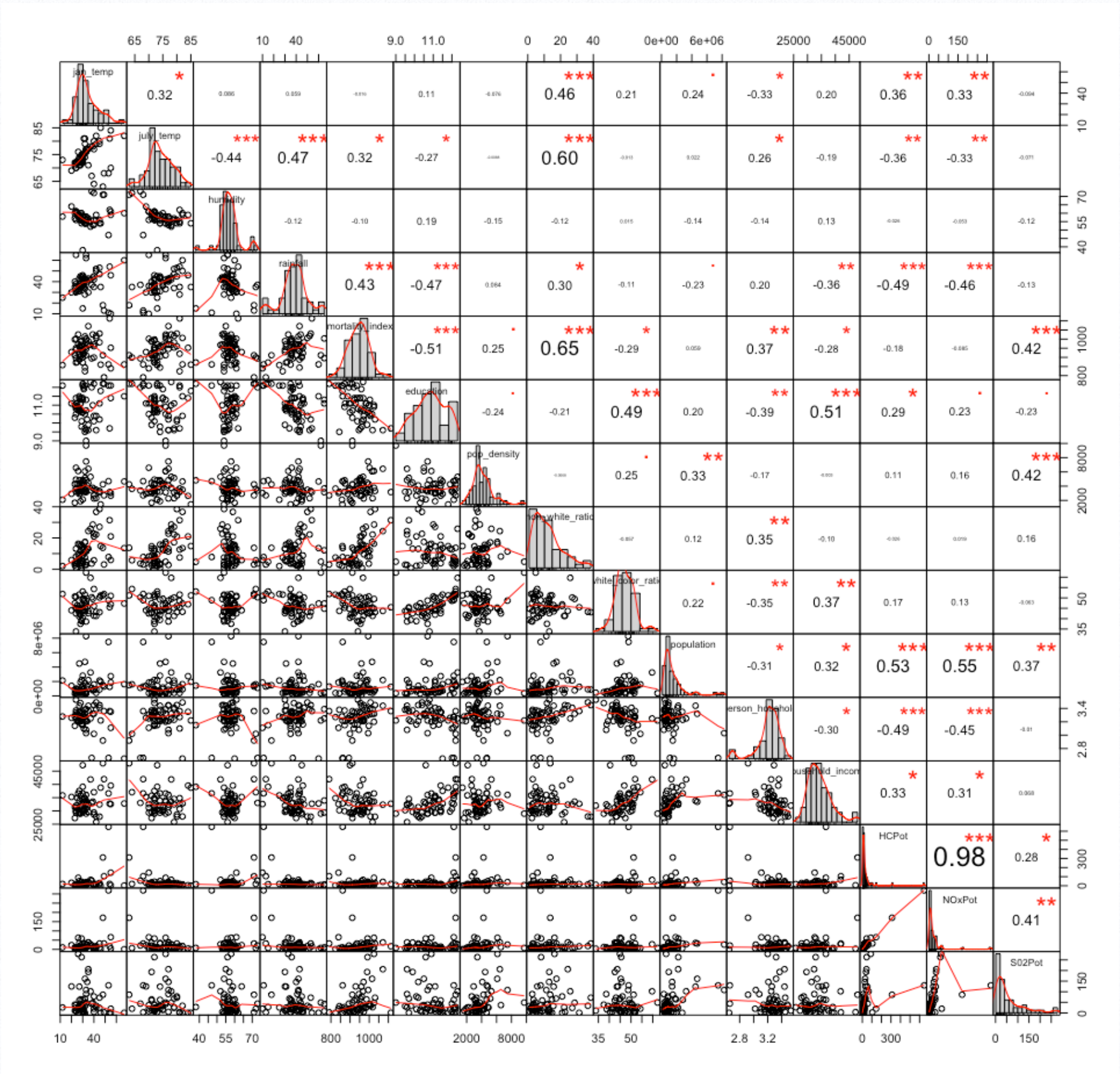
http://203.247.53.31/Stat_Notes/example_data/SMSA_USA.csv

```
smsa<-read.csv('http://203.247.53.31/Stat_Notes/example_data/SMSA_USA.csv')
names(smsa)
```

```
> names(smsa)
 [1] "city_name"          "jan_temp"           "july_temp"
 [6] "mortality_index"    "education"          "pop_density"
[11] "population"         "person_houshold"    "household_income"
[16] "S02Pot"             "southern"
```

```
# Matrix of Scatter plot
install.packages('PerformanceAnalytics') ; library(PerformanceAnalytics)
chart.Correlation(smsa[,2:16], histogram=TRUE, pch=19)
```

- 산점도에 나타난 선은 두 변수의 함수관계 : LOWESS (locally weighted scatterplot smoothing) <=> Local Regression - 직선 회귀추정식(OLS 최소자승추정값)과는 다름
- 히스토그램의 붉은 함수는 확률밀도함수임


jan_temp
july_temp
humidity
rainfall
mortality_index
education
pop_density
white_ratio
white_color_rati
population
HCPot
NOxPot
S02Pot

```
#Corelation coefficeinet
smsa.cor<-cor(smsa[,2:16], method="pearson",use="complete.obs") # "kendall", "spearman"
smsa.cor
```

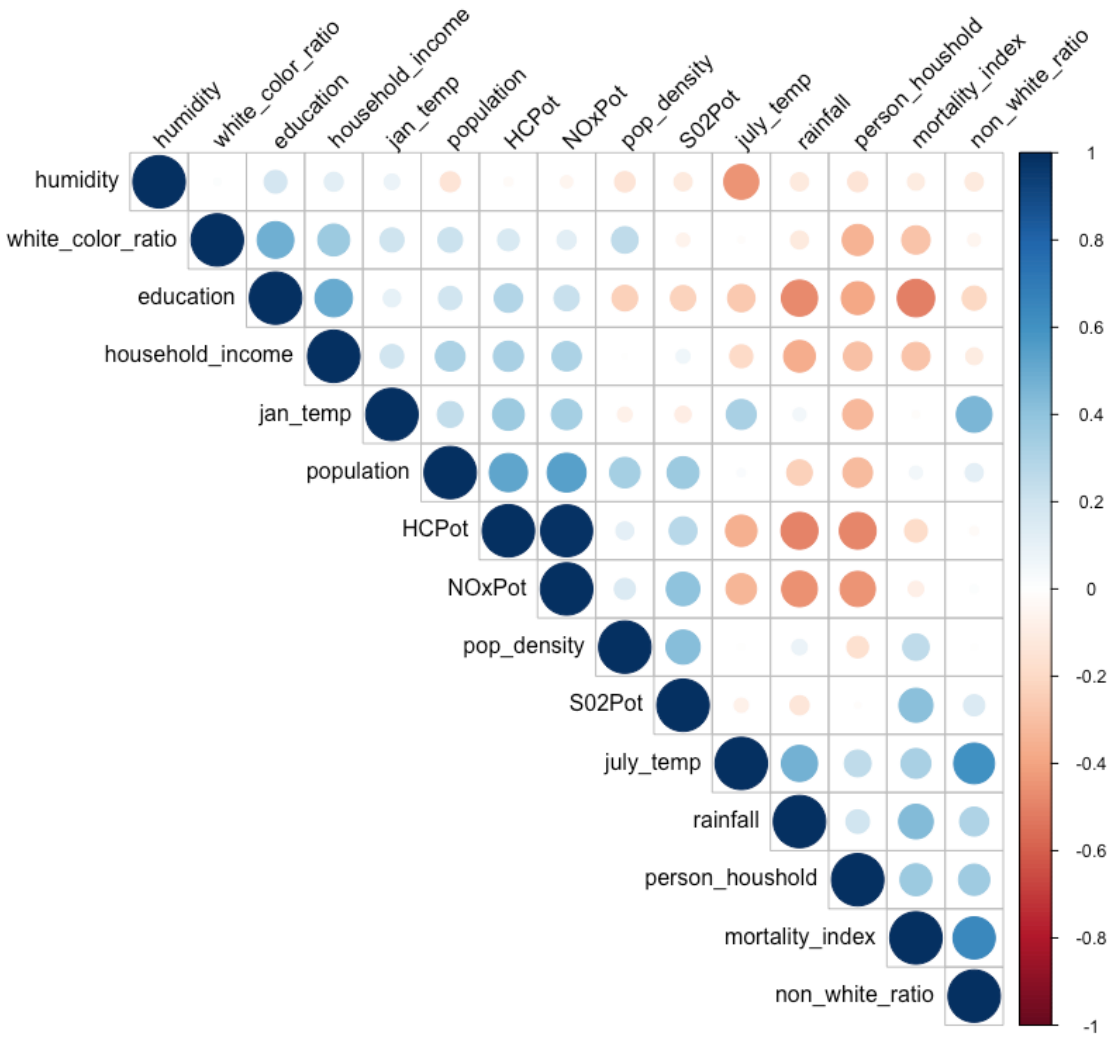
```
> smsa.cor
                  jan_temp    july_temp    humidity    rainfall mortality_index
jan_temp         1.00000000  0.322145546  0.08552171  0.05856608     -0.01595166
july_temp        0.32214555  1.000000000 -0.44139661  0.47225673      0.32182798
humidity         0.08552171 -0.441396609  1.00000000 -0.11777277     -0.10107444
rainfall         0.05856608  0.472256726 -0.11777277  1.00000000      0.43311419
mortality_index -0.01595166  0.321827975 -0.10107444  0.43311419      1.00000000
education        0.10819374 -0.269484113  0.18567008 -0.47297806     -0.50808670
pop_density     -0.07600591 -0.008832798 -0.14940354  0.08388594      0.25212106
non_white_ratio  0.45922415  0.602237473 -0.11935954  0.30276507      0.64655608
```

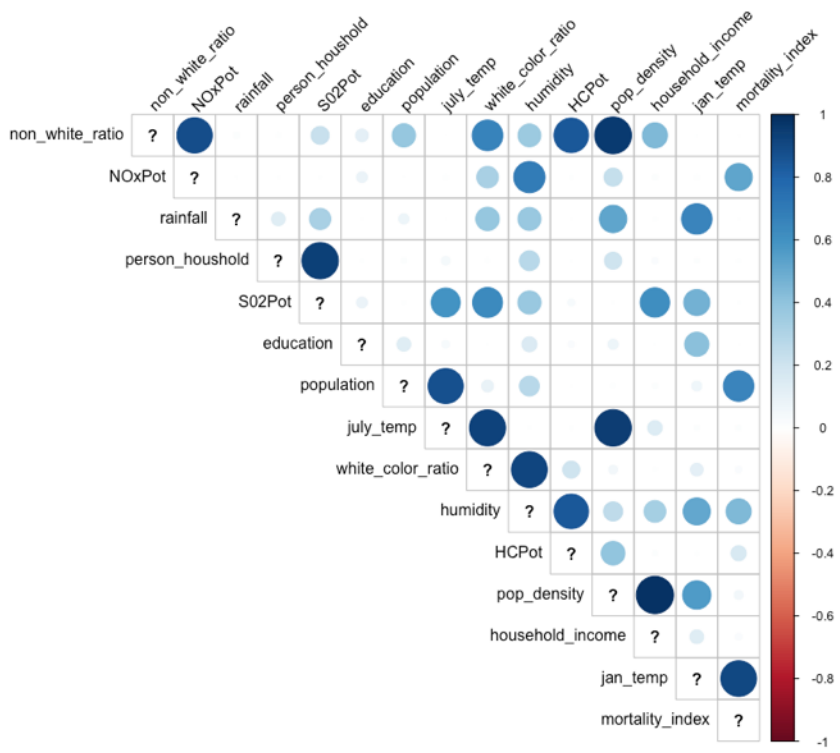
```
install.packages('Hmisc') ; library(Hmisc)
smsa.cor2<-rcorr(as.matrix(smsa[,2:16]), type="pearson")
smsa.cor2$r; smsa.cor2$P
install.packages('corrplot'); library(corrplot)
corrplot(smsa.cor2$r, type = "upper", order = "hclust", tl.col = "black", tl.srt = 45)
corrplot(smsa.cor2$P, type = "upper", order = "hclust",  tl.col = "black", tl.srt = 45)
```

• 상관계수 값 / 유의확률 출력

```
> smsa.cor2$r
                  jan_temp    july_temp    humidity
jan_temp         1.00000000  0.322145546  0.08552171
july_temp        0.32214555  1.000000000 -0.44139661
humidity         0.08552171 -0.441396609  1.00000000
```

```
> smsa.cor2$P
                     jan_temp    july_temp     humidity     rainfall
jan_temp                   NA 1.283821e-02 0.5195552815 6.594968e-01
july_temp        0.0128382052           NA 0.0004662175 1.591345e-04
humidity         0.5195552815 4.662175e-04           NA 3.743410e-01
rainfall         0.6594968054 1.591345e-04 0.3743409857           NA
mortality_index  0.9045520355 1.293209e-02 0.4462331198 6.117124e-04
education        0.4146895019 3.901654e-02 0.1591546836 1.549941e-04
pop_density      0.5672205147 9.470628e-01 0.2587337342 5.276034e-01
```


humidity
white_color_ratio
education
household_income
jan_temp
population
HCPot
NOxPot
pop_density
S02Pot
july_temp
rainfall
person_houshold
mortality_index
non_white_ratio
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

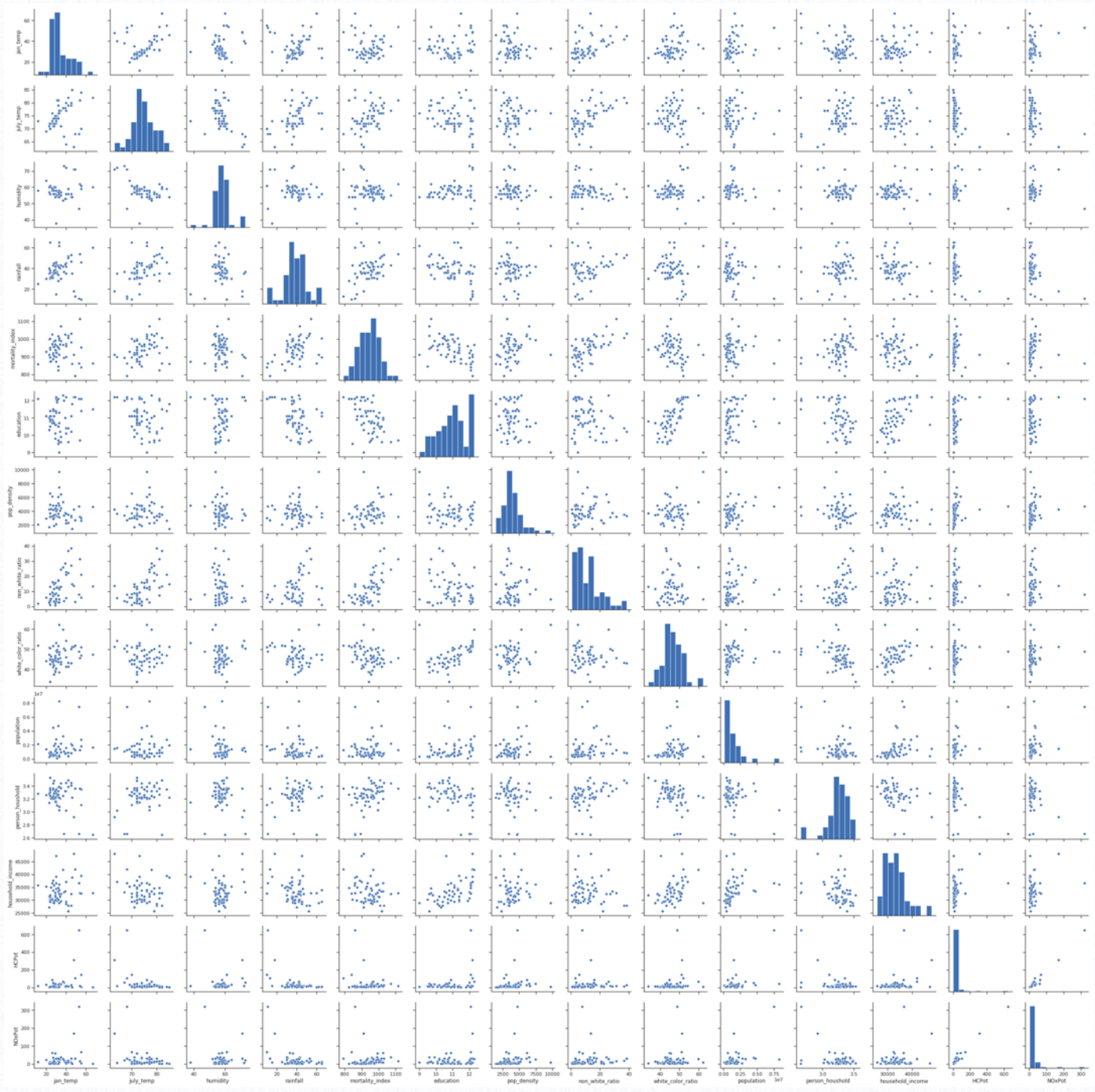

non_white_ratio
NOxPot
rainfall
person_houshold
S02Pot
education
population
july_temp
white_color_ratio
humidity
HCPot
pop_density
household_income
jan_temp
mortality_index
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

파이썬

```
import pandas as pd
smsa=pd.read_csv('http://203.247.53.31/Stat_Notes/example_data/SMSA_USA.csv')
smsa.columns
```

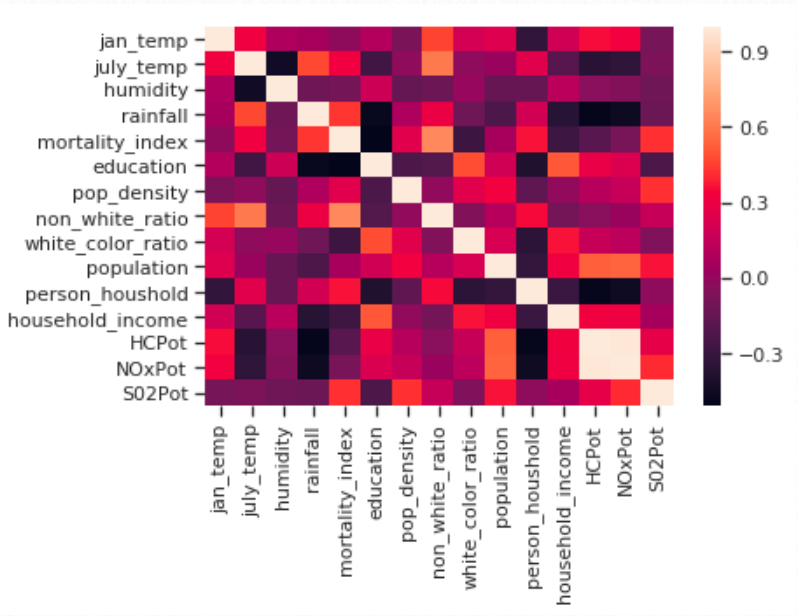
```
Index(['city_name', 'jan_temp', 'july_temp', 'humidity', 'rainfall',
       'mortality_index', 'education', 'pop_density', 'non_white_ratio',
       'white_color_ratio', 'population', 'person_houshold',
       'household_income', 'HCPot', 'NOxPot', 'SO2Pot', 'southern'],
      dtype='object')
```

```
import seaborn as sns ; sns.set(style="ticks")
sns.pairplot(smsa.iloc[:,1:15])
```



```
smsa_cor=smsa.corr()
smsa_cor.style.background_gradient(cmap='coolwarm').set_precision(3)
sns.heatmap(smsa_cor)
```

	jan_temp	july_temp	humidity	rainfall	mortality_index	educati
jan_temp	1	0.322	0.0855	0.0586	-0.016	0.108
july_temp	0.322	1	-0.441	0.472	0.322	-0.269
humidity	0.0855	-0.441	1	-0.118	-0.101	0.186
rainfall	0.0586	0.472	-0.118	1	0.433	-0.473
ortality_index	-0.016	0.322	-0.101	0.433	1	-0.508
education	0.108	-0.269	0.186	-0.473	-0.508	1
pop_density	-0.076	-0.00883	-0.149	0.0839	0.252	-0.236
on_white_ratio	0.459	0.602	-0.119	0.303	0.647	-0.209
hite_color_ratio	0.208	-0.0128	0.0148	-0.114	-0.289	0.486



유의확률(p-value) 계산하기

```
import scipy.stats as ss
ss.pearsonr(smsa.iloc[:,1],smsa.iloc[:,2])
```

```
1 import scipy.stats as ss
2 ss.pearsonr(smsa.iloc[:,1],smsa.iloc[:,2])

(0.3221455458820275, 0.012838205183050927)
```