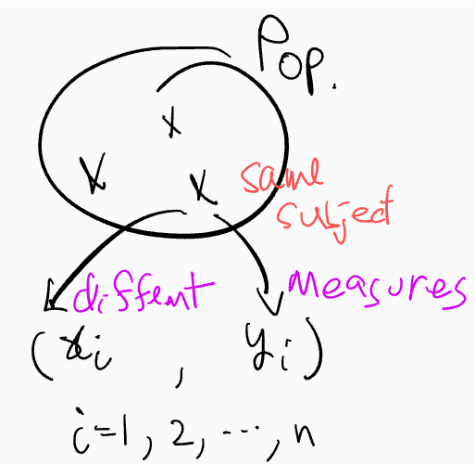
개요

상관관계?



동일 개체로부터 측정된 상이한 두 측정형 (적어도 순서형) 변수의 선형(직선)관계에 대한 척도

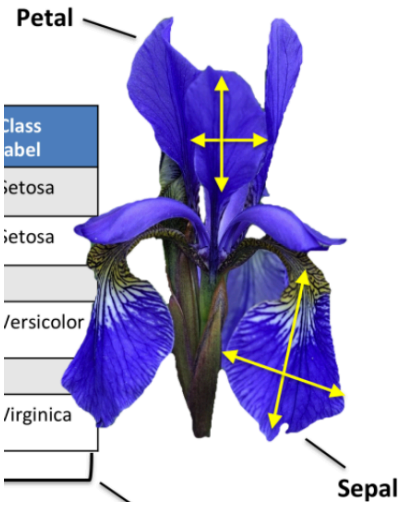
측정형(정량형) 변수간의 직선 관계 정도를 의미한다.

직선관계 정도는 상관계수 값으로 측정하면 $(-1, 1)$ 사이 값을 나타낸다.

Iris data

3개 품종 분꽃 : (Iris setosa, virginica and versicolor)

4개 변수 측정 : 꽃받침 조각(petal) 길이, 넓이 - 꽃잎(sepal) 길이, 넓이



Iris Versicolor Iris Setosa Iris Virginica

```
import pandas as pd
iris=pd.read_csv('https://vincentarelbundock.github.io/Rdatasets/csv/datasets/iris.csv')
iris.head(3)
```

	Unnamed: 0	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
0	1	5.1	3.5	1.4	0.2	setosa
1	2	4.9	3.0	1.4	0.2	setosa
2	3	4.7	3.2	1.3	0.2	setosa

- 상관관계는 4개의 측정형 변수 각각의 직선관계 정도를 나타낸다.
- 직선 상관관계라 함은 한 변수의 증감 방향과 다른 변수의 증감 방향의 일치성 혹은 역 일치성 정도를 나타내며 직선의 관계가 높을수록 두 변수의 유사성은 높다

산점도 (scatter plot)

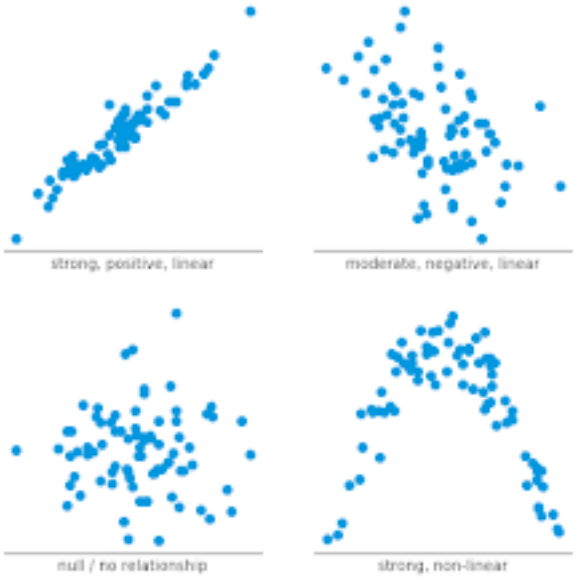
관측 데이터는 (x_i, y_i) 쌍으로 활용되어야 하며 변수간의 관계를 시각적으로 표현하는 산점도를 통하여 두 변수 간의 함수관계 $y = f(x)$ 를 보여줌

2개의 측정형 변수 데이터를 2차원 공간에 표현하여 두 변수의 함수 관계를 예상함

만약 두 변수간의 인과관계(회귀분석)를 살펴보고자 한다면,

- X-축: 결정의 요인, 설명변수, 독립변수, 예측변수
- Y-축: Output, 종속변수, 목표변수

통계 소프트웨어가 발달하면서 각 축에 분포함수까지 출력할 수 있어 변수의 정규성까지 판단할 수 있다. 정규성까지 판단이 필요한 이유는 변수 간의 관계 분석 시 정규성을 만족해야 관계(함수)의 왜곡이 발생하지 않는다.



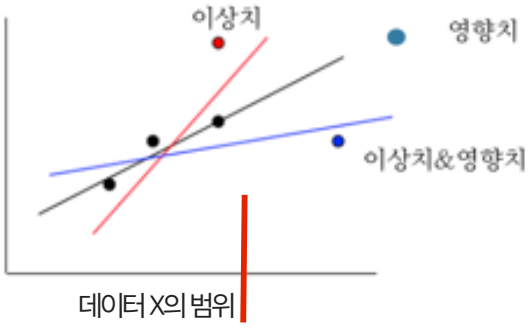
두 변수의 함수관계

- 두 변수의 함수 관계(직선 함수 관계)를 판단
- 선형 모형에서는 직선(선형)의 관계를 판단한다.
- 각 변수의 확률분포 함수 형태 파악 -> 정규변환 필요 여부 시각적 판단 -> 검정방법은 Shapiro Wilks, Anderson Darling 방법 사용 [http://203.247.53.31/Fall_2020/lecturenote/sm_good_of_fits.pdf]

이상치 영향치 진단 [선형회귀분석]

이상치 outlier

- 선형 함수 관계에서 적합 직선을 많이 벗어난 관측값 - 실제 오차의 분산 기준 2* σ를 벗어남, 설명변수 값은 관측 값의 범위 내에 있음 - 회귀계수 변동, 결정계수 낮춤



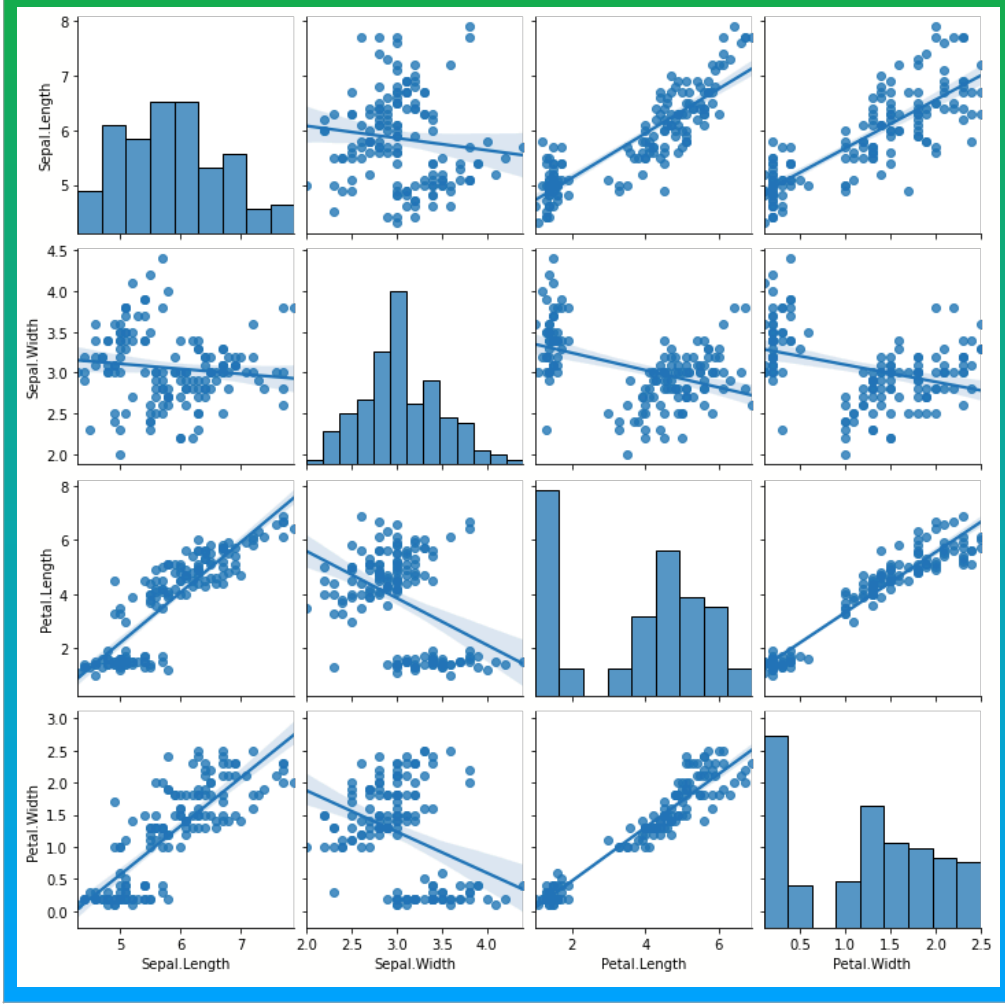
영향치 influential

- 설명변수 값이 극단 값(다른 관측치와 떨어져 있고 두 변수의 함수 관계에 영향을 주는 관측값
- 순수 영향치: 함수 회귀 추정 식 상에 있어 함수 관계(기울기 변동)에는 영향을 주지 않으나 결정계수 높여 설명변수의 설명능력(결정계수 크기)을 과다하게 높은 것으로 판단하게 하는 결과 왜곡

예제 데이터 산점도

```
import seaborn as sns
df=iris.iloc[:,1:5]
ax=sns.pairplot(df,kind="reg",corner=True,diag_kind="hist")
```

- 꽃받침 조각 넓이(sepal width)와 직선 관계가 가장 높은 변수는 꽃잎(petal) 길이이다.
- 4개 변수 중 직선 관계가 가장 강한 변수는 꽃잎 길이와 넓이이다.
- 꽃잎의 경우 길이와 넓이는 음의 직선(상관) 관계가 존재한다. 길이가 커지면 넓이는 작아진다. 반면, 꽃받침 길이는 넓이와 약한 음의 직선관계가 존재하나 꽃잎 길이와 넓이는 다소 강한 직선 관계가 존재한다.
- **[주의]** 상관관계 정도와 기울기 크기는 관계없다. 상관관계가 높다 <=> 직선 가까이 점들이 놓여 있음
- **[히스토그램]** 꽃받침 조각 길이 넓이 좌우 대칭 분포를 가짐 (정규분포 특성) 그러나 꽃잎 길이와 넓이는 작은 값들이 많아 우로 치우침 형태처럼 보이나 작은 값을 제외하면 좌우 대칭이다. => 정규변환이 필요한가?



```
import scipy.stats as stats
data=iris['Sepal.Length']
stats.shapiro(data),stats.anderson(data,'norm')
```

꽃잎 길이 : 유의확률 1%로 귀무가설 기각 – 정규분포 따르지 않음 (좌우 대칭 오케이)

```
((0.9760897755622864, 0.01017984002828598),
 AndersonResult(statistic=0.8891994860134105, critical_values=array([0.562, 0.64 , 0.767, 0.895, 1.065]),
                significance_level=array([15. , 10. ,  5. ,  2.5,  1. ])))
```

```
import scipy.stats as stats
import numpy as np
data=iris['Petal.Width']
stats.shapiro(np.sqrt(data)),stats.anderson(np.sqrt(data),'norm')
```

우로 치우친 형태를 가지고 있어 제곱근 변환에 대한 정규성 검정 결과 유의확률은 더 작아져 정규분포 향태로부터 더 멀어졌다. 이유는 작은 값이 많은 것이지 우로 치우친 향태는 아님을 알 수 있다. 이런 경우 정규변환을 적용하기보다는 원 데이터 값을 그대로 사용하여 향후 분석하는 것이 적절하다.

```
((0.82148677110672, 3.0511385518822154e-12),
 AndersonResult(statistic=11.561987140402096, critical_values=array([0.562, 0.64 , 0.767, 0.895, 1.065]),
```

꽃받침 조각 길이 : 귀무가설 기각 – 정규분포 따르지 않음

```
((0.8762688040733337, 7.412849778454245e-10),
 AndersonResult(statistic=7.6785455198266845, critical_values=array([0.562, 0.64 , 0.767, 0.895, 1.065]),
```

꽃잎 넓이 : 귀무가설 기각 – 정규분포 따르지 않음

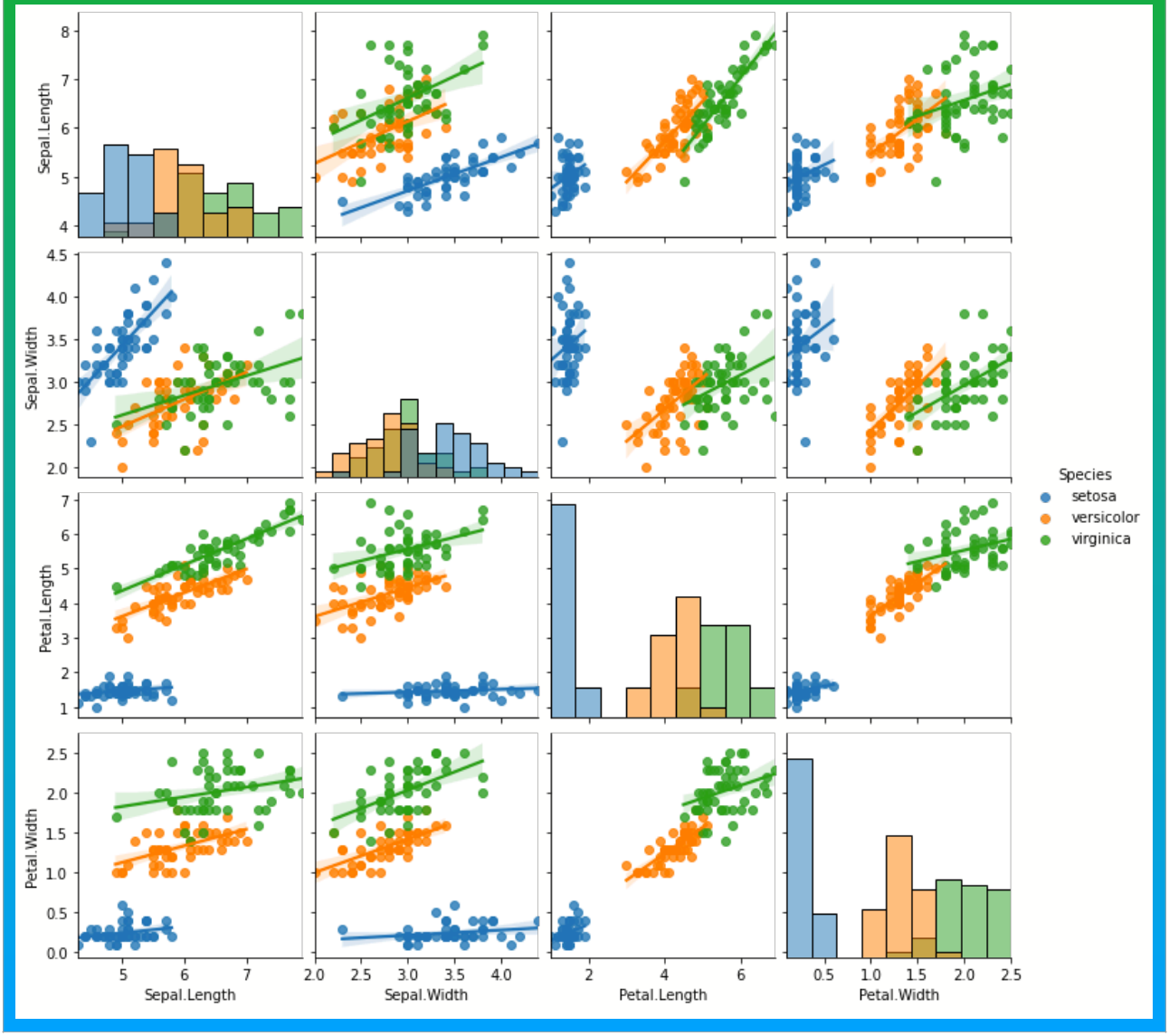
```
((0.9018339514732361, 1.6802413682626138e-08),
 AndersonResult(statistic=5.1056620354169695, critical_values=array([0.562, 0.64 , 0.767, 0.895, 1.065]),
```

꽃받침 조각 넓이 : 귀무가설 채택 – 정규분포 따름

```
((0.9849170446395874, 0.10113201290369034),
 AndersonResult(statistic=0.9079550471145126, critical_values=array([0.562, 0.64 , 0.767, 0.895, 1.065]),
```

```
import seaborn as sns
df=iris.iloc[:,1:6]
ax=sns.pairplot(df,hue='Species',kind="reg",diag_kind="hist")
```

- 3개 품종 : setosa, virginica and versicolor 4개 측정변수 산점도를 그린 것이다.
- Petal length 꽃잎 길이의 경우 setosa 품종의 길이가 다른 품종에 비해 짧음을 알 수 있다. 앞에서 꽃잎 길이가 우로 치우친 형태를 가진 이유가 setosa 품종때문임을 알 수 있다.
- 꽃잎 길이와 넓이 산점도(첫 행 두번째 열)를 보면 (1) 모든 품종에서 두 변수는 양의 상관관계(길이가 커지면 넓이가 커진다)를 가지며 (2) 상관(직선) 관계 정도가 높으며(점들이 직선 상에 가까이 놓여 있음) (3) 두 변수의 상관관계가 가장 높은 품종은 setosa{다른 품종에 비해 직선에 가장 가까이 놓였음} (4) 넓이 변화에 따른 길이 변화량(기울기)이 가장 큰 품종은 versicolor이다.



상관계수 계산

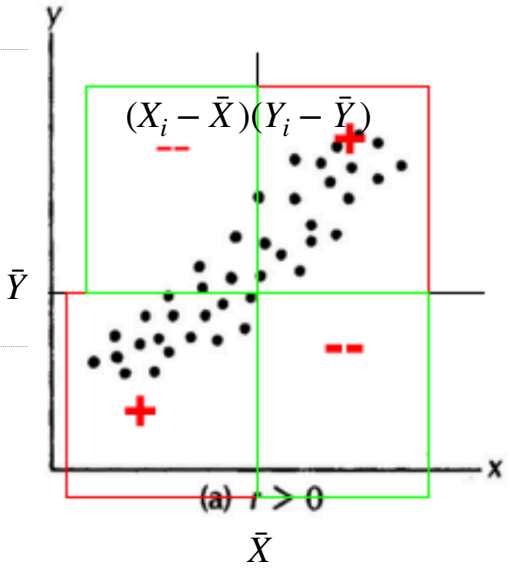
피어슨 Pearson 상관계수

측정형 변수 간의 선형관계 척도

모집단 상관계수

$$\rho = \frac{COV(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}},$$

$$COV(X,Y) = E(X - E(X))(Y - E(Y))$$



표본

표본 공분산 : $COV(X,Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$

표본 피어슨 상관계수 $r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$

- 공분산은 단위를 가지지만 상관계수는 단위의 표준화로 절대단위의 값이다.
- 분모는 확률변수의 표준편차이므로 항상 양이므로 상관계수의 부호는 분자항(공분산)에 의해 결정된다.
- $(x_i - \bar{x})(y_i - \bar{y})$의 부호는 그림(수평선은 Y의 평균, 수직선은 X의 평균, 오른쪽의 관측치 5개를 제외한 경우)에서 시각적으로 확인할 수 있음.

```
df=iris.iloc[:,1:5]
df.corr(method='pearson') #default
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.000000	-0.117570	0.871754	0.817941
Sepal.Width	-0.117570	1.000000	-0.428440	-0.366126
Petal.Length	0.871754	-0.428440	1.000000	0.962865
Petal.Width	0.817941	-0.366126	0.962865	1.000000

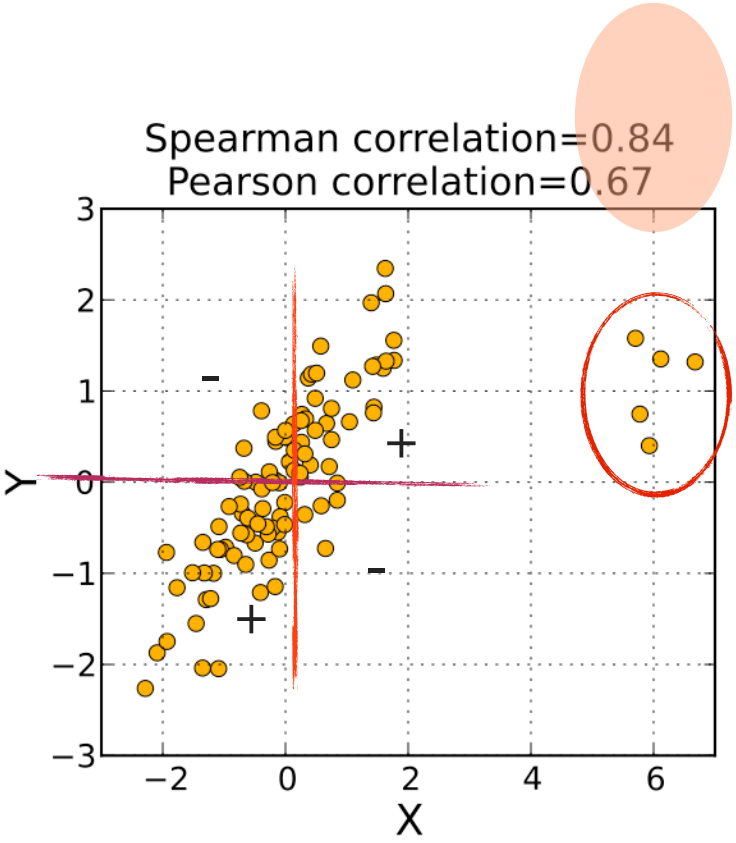
스피어맨 Spearman 순위 상관계수

순서형 변수(측정형 변수도 순위 변환 후) 간의 선형 관계 척도

(계산식 1) $r_s = Corr(R_{X_i}, R_{Y_i})$ where R_{X_i}는 X_i의 크기 순위이며, R_{Y_i}는 Y_i의 크기 순위이다.

(계산식 2) $r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$, $d_i = R_{X_i} - R_{Y_i}$

[피어슨, 스피어만 상관계수 산점도를 보면] 대부분의 데이터 범위 밖에 있는 관측치(타원형 내 관측치)는 상관계수 값을 높이는 역할을 한다. 그러므로 상관계수를 계산하기 전에 반드시 산점도를 그려 데이터의 범위를 많이 벗어난 관측치가 있는지 확인하여 상관분석의 활용도를 높일 필요가 있음.
순위 상관계수는 순위에 의한 상관(직선) 관계 척도이므로 이상치에 영향을 받지 않아 피어슨 상관관계보다 낮아진다. 만약 5개 점이 위로 이동하여 붉은 타원에 있어 직선 상에 존재한다면(이상치가 아닌 영향치가 된다면) 피어슨 상관계수가 커진다. 그러므로 상관계수 값을 구하기 전에 산점도를 그려 사전 판단을 통하여 상관계수 해석의 왜곡을 없애야 한다.




```
df=iris.iloc[:,1:5]
df.corr(method='spearman')
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.000000	-0.166778	0.881898	0.834289
Sepal.Width	-0.166778	1.000000	-0.309635	-0.289032
Petal.Length	0.881898	-0.309635	1.000000	0.937667
Petal.Width	0.834289	-0.289032	0.937667	1.000000

Kendall Tau 순위 상관계수

순서형 변수 간의 선형 관계 척도로 concordant (쌍의 관측치 값의 크기와 순위의 크기가 일치하는 정도로 판단함)

(계산식) $\tau = \frac{\#of\_concordant\_pairs - \#of\_disconcordant\_pairs}{n(n-1)/2}$

- 만약 $(x_i > x_j), (y_i > y_j)$ 이거나 $(x_i < x_j)$ 이면, $(y_i < y_j)$ 이면 두 관측치는 concordant 쌍이라 함
- τ값이 클수록 데이터 순위의 일치도는 높아지므로 상관관계가 높아진다.

```
df=iris.iloc[:,1:5]
df.corr(method='kendall')
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.000000	-0.076997	0.718516	0.655309
Sepal.Width	-0.076997	1.000000	-0.185994	-0.157126
Petal.Length	0.718516	-0.185994	1.000000	0.806891
Petal.Width	0.655309	-0.157126	0.806891	1.000000

Pandas corr() 결과는 데이터 프레임으로 저장되고 행 인덱스, 열 변수명은 모두 원 측정변수 이름이다.

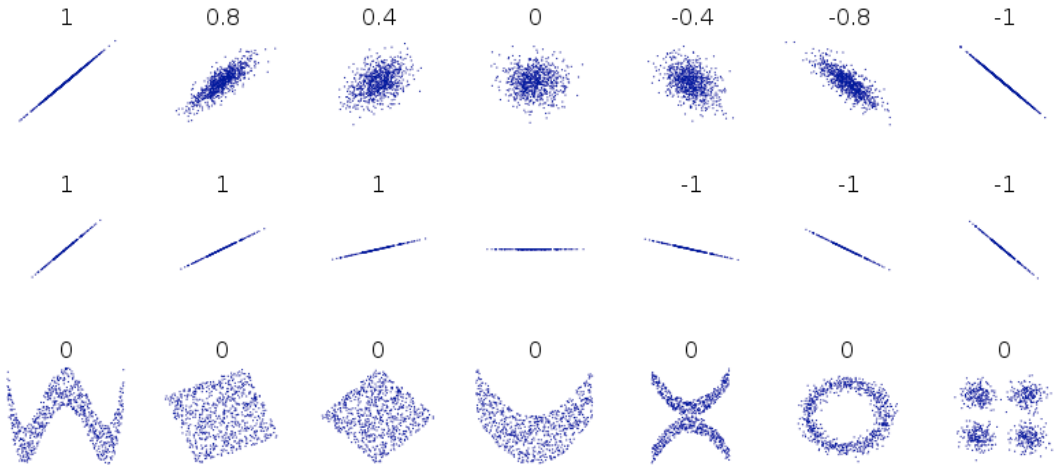
```
df=iris.iloc[:,1:5]
df_cor=df.corr(method='pearson') #default
type(df_cor), df_cor.columns, df_cor.index
```

```
(pandas.core.frame.DataFrame,
 Index(['Sepal.Length', 'Sepal.Width', 'Petal.Length', 'Petal.Width'],
 Index(['Sepal.Length', 'Sepal.Width', 'Petal.Length', 'Petal.Width'],
```

상관계수 해석

다음 판단 기준은 측정형 데이터, 표본크기가 20개 이상(동일 값의 관측값 거의 없다는 가정) 인 경우 적용 가능하다.

- -1과 1사이의 값이다.
- 1에 가까우면 양의 선형 상관 관계가 존재한다. 한 변수의 값이 증가(감소)하면 다른 변수 값도 증가(감소)한다.
- -1에 가까우면 음의 선형 상관 관계가 존재한다. 한 변수의 값이 증가(감소)하면 다른 변수 값은 감소(증가)한다.
- 두 변수의 상관 관계가 높다는 것은 두 변수가 동일한(comparable) 개념을 측정한다는 의미도 담고 있다(두 변수가 유사함). 그러므로 변수를 축약하거나 개체를 분류하는데 사용되는 다변량 분석에서는 공분산, 혹은 상관계수 개념 사용
- 상관 계수가 0에 가깝다는 것은 선형 상관 관계가 없다는 것이지 함수 관계가 없다는 것은 아니다. 아래 그림에서 두 변수는 함수관계(아치식, 4차식, 원 모양 등)는 존재하지만 상관계수는 0이다.
- 강조하지만 상관계수는 두 측정형 변수의 직선 함수 관계만에 대한 척도이다.



Rule of thumb:

- $0.0 = |r|$: no correlation
- $0.0 < |r| < 0.2$: very weak correlation
- $0.2 \leq |r| < 0.4$: weak correlation
- $0.4 \leq |r| < 0.6$: moderately strong correlation
- $0.6 \leq |r| \leq 0.8$: strong correlation
- $0.8 \leq |r| < 1.0$: very strong correlation
- $1.0 = |r|$: perfect correlation

: 상관계수의 제곱이 회귀분석의 결정계수이므로 70% 이상 되어야 회귀모형(예측모형)의 정도가 높다고 할 수 있다.

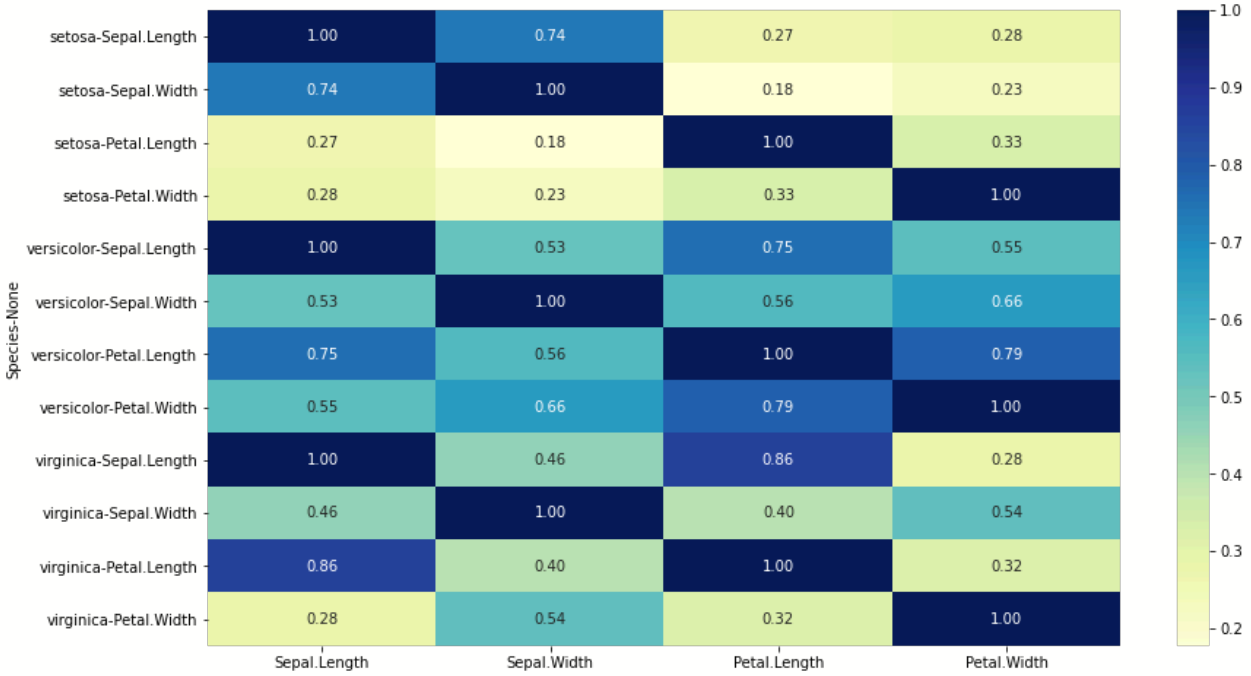
집단간 측정변수 상관계수 구하기

```
df=iris.iloc[:,1:6]
df_cor=df.groupby('Species').corr()
df_cor.style.background_gradient(cmap='coolwarm').set_precision(3)
```

Species		Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
setosa	Sepal.Length	1.000	0.743	0.267	0.278
	Sepal.Width	0.743	1.000	0.178	0.233
	Petal.Length	0.267	0.178	1.000	0.332
	Petal.Width	0.278	0.233	0.332	1.000
versicolor	Sepal.Length	1.000	0.526	0.754	0.546
	Sepal.Width	0.526	1.000	0.561	0.664
	Petal.Length	0.754	0.561	1.000	0.787
	Petal.Width	0.546	0.664	0.787	1.000
virginica	Sepal.Length	1.000	0.457	0.864	0.281
	Sepal.Width	0.457	1.000	0.401	0.538
	Petal.Length	0.864	0.401	1.000	0.322
	Petal.Width	0.281	0.538	0.322	1.000

```
import seaborn as sn
import matplotlib.pyplot as plt
plt.rcParams["figure.figsize"] = (14,8)
sn.heatmap(df_cor,annot=True,fmt=".2f",cmap="YlGnBu")
plt.show()
```

상관계수 0.8 이상 : virginica 꽃잎 길이, 꽃받침 조각 길이 0.86 가장 높음



상관계수 추론

1집단 상관계수 유의성 검정

가설

- 귀무가설 : 두 변수의 **직선** 상관관계는 유의하지 않다. <=> 서로 독립이다. $H_0 : \rho = 0$
- 대립가설 : 두 변수의 **직선** 상관관계는 유의하다. $H_1 : \rho \neq 0$

데이터 검증

- 데이터는 이변량 정규분포에 근사해야 한다. 단 n>20 인 대표본에서는 문제 없음
- 산점도를 그려 데이터 범위(X-) 밖의 관측치 존재 여부를 체크한다. - 존재한다면 제외하거나 활용 시 주의해야 한다.

검정통계량 ~ t분포

$$TS = \frac{r}{\sqrt{\frac{(1-r^2)}{(n-2)}}} \sim t(n-2), n = \text{표본크기}, r = \text{표본상관계수}$$

해석

유의확률($P(|TS| > t(1-\alpha/2; n-2))$)이 유의수준보다 작다면 귀무가설을 기각하여 상관관계의 유의하다고 결론내리고 표본상관계수의 부호를 이용하여 해석

- 귀무가설이 기각, 표본상관계수 부호 + => 두 변수는 양의 상관관계가 있고 한 변수의 값이 증가(감소)하면 다른 변수의 값도 증가(감소)한다.
- 귀무가설이 기각, 표본상관계수 부호 - => 두 변수는 음의 상관관계가 있고 한 변수의 값이 증가(감소)하면 다른 변수의 값도 감소(증가)한다.

꽃잎 길이와 넓이 상관계수 0.96(양의 상관관계), 매우 유의함 (p<0.001)

```
import scipy.stats as st
stats.pearsonr(iris['Petal.Length'],iris['Petal.Width'])
```

```
(0.962865431402796, 4.6750039073285846e-86)
```

꽃받침 조각 길이와 넓이 상관계수 -0.11(음의 상관관계), 유의하지 않음 (p=0.15)

```
stats.pearsonr(iris['Sepal.Length'],iris['Sepal.Width'])
```

```
(-0.11756978413300206, 0.15189826071144766)
```

회귀계수와 관계 $Y = \alpha + \beta X + e$

독립변수 X가 Y에 선형적 영향을 미치는지 검정 <=> 기울기 $H_0 : \beta = 0$(영향을 미치는 않음) 유의성 검정 <=> 상관계수의 유의성 검정($H_0 : \rho = 0$)과 동일하다.

$$\hat{\beta} = \sqrt{\frac{S_{XY}}{S_{XX}}} r, S_{XX} = \sum (X_i - \bar{X})^2, S_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y})$$

- 상관계수 부호와 회귀계수 부호는 동일하다.
- 두 통계량 모두 $t(n-2)$를 검정통계량으로 갖는다. 그러므로 두 통계량의 유의확률은 동일하다.
- 단순 회귀모형에서 결정계수 Determination Coefficient($R^2 = \frac{SST}{SSR} = \frac{\sum (y_i - \bar{y})^2}{\sum (\hat{y}_i - \bar{y})^2}$)의 제곱근은 상관계수이다. $r = \pm \sqrt{R^2}$

$PW = -0.36 + 0.42 * PL$, 상관계수=0.96
유의확률 동일(상관계수, 회귀계수 기울기)

```
from scipy import stats
stats.linregress(iris['Petal.Length'],iris['Petal.Width'])
```

```
LinregressResult(slope=0.41575541635241114, intercept=-0.36307552131902776
rvalue=0.9628654314027963, pvalue=4.6750039073255014e-86, stderr=0.009582435790
```

```
import statsmodels.api as sm
model=sm.OLS(iris['Petal.Width'],sm.add_constant(iris['Petal.Leng
th']))
results=model.fit()
results.summary()
```

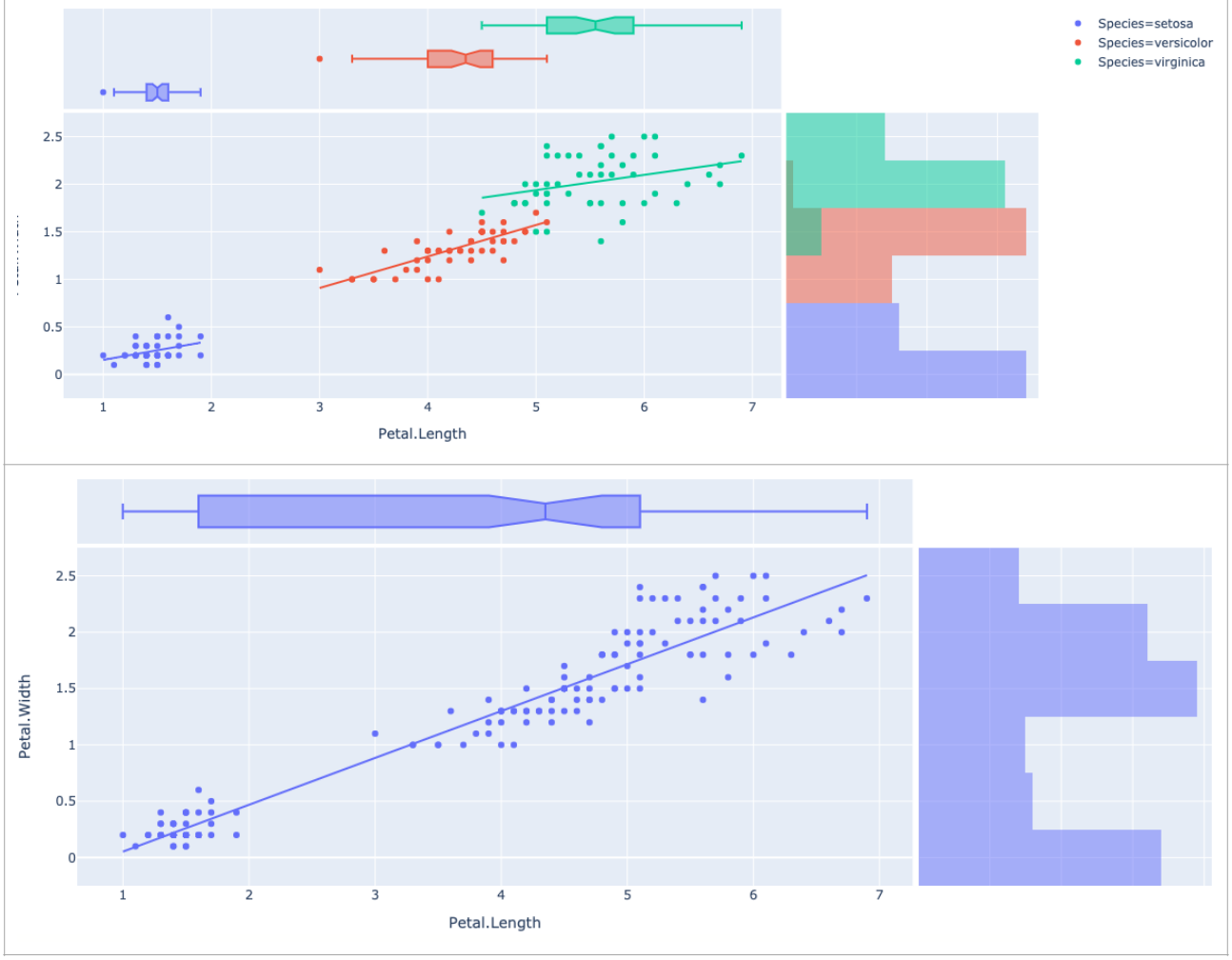
OLS Regression Results

Dep. Variable:	Petal.Width	**R-squared:**	0.927
Model:	OLS	**Adj. R-squared:**	0.927
Method:	Least Squares	**F-statistic:**	1882.
Date:	Sat, 17 Oct 2020	**Prob (F-statistic):**	4.68e-8
Time:	23:28:33	**Log-Likelihood:**	24.796
lo. Observations:	150	**AIC:**	-45.59
Df Residuals:	148	**BIC:**	-39.57
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.3631	0.040	-9.131	0.000	-0.442	-0.285
Petal.Length	0.4158	0.010	43.387	0.000	0.397	0.435

집단간 측정변수 상관계수 구하기

```
from matplotlib import pyplot as plt
import plotly.express as px
fig=px.scatter(iris,y='Petal.Width',x='Petal.Length',color='Species',marginal_x='box',marginal_y='histogram', trendline='ols')
fig.show()
```



특별한 경우

1집단 상관계수 일정값 갖는지 검정 : 귀무가설 : $H_0 : \rho = \rho_0$

• 활용 : 미국 대학생의 경우 공부시간과 학점의 상관계수는 0.7이었다. 그럼 한국 대학생 상관계수는 0.7인가?

• 검정통계량 : $TS = \frac{1}{2} ln(\frac{1+r}{1-r}) \sim N(\frac{1}{2} ln(\frac{1+\rho_0}{1-\rho_0}), \frac{1}{(n-3)})$

한국산 분꽃의 경우 꽃받침(sepal) 길이와 꽃잎(petal) 길이의 상관계수는 0.8이었다. 피셔 예제 분꽃의 경우 한국산과 동일한 상관계수를 갖는가?

```
import pandas as pd
iris=pd.read_csv('https://vincentarelbundock.github.io/Rdatasets/csv/datasets/iris.csv')
iris.iloc[:,[1,3]].corr()
```

	Sepal.Length	Petal.Length
Sepal.Length	1.000000	0.871754
Petal.Length	0.871754	1.000000

```
import numpy as np
import scipy.stats as st
n=iris.shape[0]
rho=0.8
r=0.871754
ts=(0.5*np.log((1+r)/(1-r))-0.5*np.log((1+rho)/(1-rho)))/
np.sqrt(1/(n-3))
p_value=(1-st.norm.cdf(abs(ts),0,1))*2
ts,p_value
```

```
(2.930798464036547, 0.0033809204220280886)
```

귀무가설 기각, 피셔 분꽃의 경우 꽃받침, 꽃잎 길이 상관계수는 0.87로 한국산 분꽃보다 상관계수가 유의하게 낮다.

두 독립 모집단 상관계수 차이 검정 : 귀무가설 : $H_0 : \rho_1 = \rho_2$: 두 모집단 상관계수 크기는 동일하다.

- 대립가설 : $H_1 : \rho_1 \neq \rho_2$
- 검정통계량 : $TS = \frac{1}{2} ln(\frac{1+r_1}{1-r_1}) - \frac{1}{2} ln(\frac{1+r_2}{1-r_2}) \sim N(0, \frac{1}{(n_1-3)} + \frac{1}{(n_2-3)})$

꽃받침(sepal) 길이와 꽃잎(petal) 길이의 상관계수는 (versicolor, virginica) 품종간 동일한가?

```
df=iris.iloc[:,[1,3,5]]
df.groupby('Species').corr()

iris['Species'].value_counts()
```

```
setosa        50
virginica     50
versicolor    50
```

Species		Sepal.Length	Petal.Length
setosa	Sepal.Length	1.000000	0.267176
	Petal.Length	0.267176	1.000000
versicolor	Sepal.Length	1.000000	0.754049
	Petal.Length	0.754049	1.000000
virginica	Sepal.Length	1.000000	0.864225
	Petal.Length	0.864225	1.000000

```
import numpy as np
import scipy.stats as st
n1=50
n2=50
r1=0.754049
r2=0.864225
ts=(0.5*np.log((1+r1)/(1-r1))-0.5*np.log((1+r2)/(1-r2)))/
np.sqrt(1/(n1-3)+1/(n2-3))
p_value=(1-st.norm.cdf(abs(ts),0,1))*2
ts,p_value
```

유의확률이 11.2%로 귀무가설을 기각하지 못한다. (versicolor, virginica) 품종간 상관계수의 차이는 유의하지 않다.

```
(-1.5877407387688913, 0.11234497687180856)
```

상관계수 활용

변수의 유사성 척도

상관계수는 두 변수간의 직선 관계에 대한 척도이다. 상관관계가 높다는 것은(상관계수가 ± 1에 가깝다) 두 변수가 가진 정보의 중첩도가 높다는 것이고 관심 모집단 개체에 대한 정보를 얻는데 두 변수 모두가 필요하지는 않다는 것이다.

예를 들어 키와 몸무게는 상관관계가 높으므로 둘 다 측정하여 신체 특성을 파악할 필요는 없다. 기성복 하의를 구매할 때도 길이(허리에서 무릎, 무릎에서 발목)에 대한 정보는 기장에 넓이(허리둘레, 허벅지 둘레, 발목 둘레)에 대한 정보는 허리둘레와 상관관계가 높아 2개의 측정변수만으로 구매 가능하다.

이처럼 상관관계는 변수의 유사성에 대한 척도로 높은 변수들은 일부만 사용해도 된다.

회귀분석

종속변수 설명력 높은 변수 사전 선택

설명변수(열의 크기)가 많고 데이터 개수(행의 크기)가 적은 경우 과적합문제 등이 발생하므로 사전에 일정 크기만큼 설명변수의 수를 줄이는 것이 적절하다. 하여, 빅데이터에서는 종속변수와 설명변수들간의 상관계수를 활용하여 일정 기준 이상의 설명변수만으로 예측모형을 도출한다.

미국 도시 사회지표 (사망률지수 mortality index : Y)

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 59 entries, 0 to 58
Data columns (total 17 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   city_name          59 non-null     object 
 1   jan_temp           59 non-null     int64  
 2   july_temp          59 non-null     int64  
 3   humidity           59 non-null     int64  
 4   rainfall           59 non-null     int64  
 5   mortality_index    59 non-null     float64
 6   education          59 non-null     float64
 7   pop_density        59 non-null     int64  
 8   non_white_ratio    59 non-null     float64
 9   white_color_ratio  59 non-null     float64
 10  population         59 non-null     int64  
 11  person_houshold    59 non-null     float64
 12  household_income   59 non-null     int64  
 13  HCPot              59 non-null     int64  
 14  NOxPot             59 non-null     int64  
 15  SO2Pot             59 non-null     int64  
 16  southern           59 non-null     object 
```

미국 도시 사회지표_사망률 지수 상관계수 계산

```
df=smsa.iloc[:,1:15]
df_cor=df.corr()
df_cor
```

	jan_temp	july_temp	humidity	rainfall	mortality_index
jan_temp	1.000000	0.322146	0.085522	0.058566	-0.015952
july_temp	0.322146	1.000000	-0.441397	0.472257	0.321828
humidity	0.085522	-0.441397	1.000000	-0.117773	-0.101074

사망률 지수 mortality index 상관계수가 0.3 이상인 설명변수만 선택하여 리스트에 저장했음

```
select_feature=df_cor[abs(df_cor['mortality_index'])>0.3].index.tolist()
select_feature
```

```
['july_temp',
 'rainfall',
 'mortality_index',
 'education',
 'non_white_ratio',
 'person_houshold']
```

```
df_reg=df[select_feature] #최종 분석대상 데이터
df_reg.corr()
```

	july_temp	rainfall	mortality_index	education	non_white_ratio	person_houshold
july_temp	1.000000	0.472257	0.321828	-0.269484	0.602237	0.257080
rainfall	0.472257	1.000000	0.433114	-0.472978	0.302765	0.199056
nortality_index	0.321828	0.433114	1.000000	-0.508087	0.646556	0.368016
education	-0.269484	-0.472978	-0.508087	1.000000	-0.208875	-0.389103
ıon_white_ratio	0.602237	0.302765	0.646556	-0.208875	1.000000	0.352736
erson_houshold	0.257080	0.199056	0.368016	-0.389103	0.352736	1.000000

다중공선성 문제 해결 [http://203.247.53.31/2020spring/LM2020/LinearModel_2020_spring_ch4.pdf]

회귀모형 $y = X\beta + e$의 β의 OLS 추정치는 $\beta = (X'X)^{-1}X'y$이다. 설명변수(X) 간의 상관관계가 높으면 $det(X'X) \simeq 0$이므로 $X'X^{-1}$의 값이 매우 커진다.

이로 인하여, β의 OLS 추정값이 변동이 커지고 (추정분산이 커짐) 추정계수의 부호까지 바뀌는 문제가 발생한다.

사망률 지수(Y)와 7월기온(july_temp)의 상관계수는 +0.32 양이었는데 회귀계수는 -3.6891로 음의 영향을 주는 것으로 추정되었다. 이렇게 상관계수와 회귀계수의 부호가 바뀐 것은 설명변수들간 높은 상관관계로 다중공선성이 발생하여 야기시킨 문제이다.

```
import statsmodels.api as sm
y=df_reg['mortality_index']
X=df_reg.iloc[:,[0,1,3,4,5]]
model=sm.OLS(y,sm.add_constant(X))
results=model.fit()
results.summary()
```

	coef	std err	t	P>|t|	[0.025	0.975]
const	1357.4098	192.452	7.053	0.000	971.401	1743.419
july_temp	-3.6891	1.572	-2.347	0.023	-6.842	-0.537
rainfall	1.0378	0.574	1.809	0.076	-0.113	2.189
education	-24.8046	7.581	-3.272	0.002	-40.009	-9.600
non_white_ratio	4.6503	0.769	6.046	0.000	3.108	6.193
person_houshold	10.8024	33.273	0.325	0.747	-55.934	77.539

설명변수(X)들의 높은 상관관계로 다중공선성이 발생하는 경우 설명변수(원변수)의 공분산(상관계수)행렬로부터 고유값, 고유벡터(부하, 선형계수 L)를 도출하고 이를 이용하여 얻은 주성분변수를 이용한다.
주성분변수(LX)는 원변수의 선형결합으로 계산되며 원변수와는 달리 상관계수 0인 서로 독립이다. 주성분변수를 설명변수로 종속변수를 예측(회귀모형)한다.

참고 강의노트 : http://203.247.53.31/Fall_2020/lecturenote/mva_pca.pdf