8장 과제

다음 패키지들이 사용된다.

In [1]:

```
import pandas as pd
import numpy as np
import seaborn

from scipy import stats
from matplotlib import pyplot as plt
from statsmodels.formula.api import ols
```

예제 1

handspan.txt는 3 종류의 column으로 이루어져 있다. (Sex, Height, HandSpan) 이 자료를 이용하여 회귀분석 등을 진행하고자 한다.

In [2]:

```
import pandas as pd
handspan = pd.read_table("handspan.txt", sep = '₩t')
```

1-1. 위와 같이 handspan 자료를 불러와서 HandSpan vs. Height 의 산점도를 그려보자. seaborn 혹은 pyplot.scatter을 사용하면 얻을 수 있다. 그리고 두 변수 간에 연관성이 존재한다고 할 수 있는지 간단히 기술하자.

1.2. 다음 ols 메소드를 사용해 y = Height, x = HandSpan으로 두고 단순 선형 회귀 분석을 시행하자. 결과로 나온 회귀식을 구체적으로 적어보자. (즉, Height = β_1 * HandSpan + β_0 에서 β_1과 β_0를 구체적인 수치로 적자.)

In [3]:

```
from statsmodels.formula.api import ols
```

1.3. 위에서 구한 β_1, β_0를 이용하여 선형 플롯 $y = \beta_0 + \beta_1 x$를 1.1에서 그렸던 산점도와 동시에 그려보자. 이 때 x의 구간은 [15, 30]으로 주자. 이 때 플롯을 그리는 방법은 다음을 참고하라.

HINT : 아래 코드의 임의의 점 z, w는 각각 HandSpan과 Height가 될 것이며, 직선 y = ax + b 는 1.2에서 구한 선형 식이 될 것이다.

==========

선형 플롯과 산점도를 한 그래프 안에 그리기 (임의의 7개 점 z, w를 이용해 산점도를 그리고, x의 구간을 [0, 10]으로 주고 y = 2x + 1 일차함수 그래프를 그리는 코드)

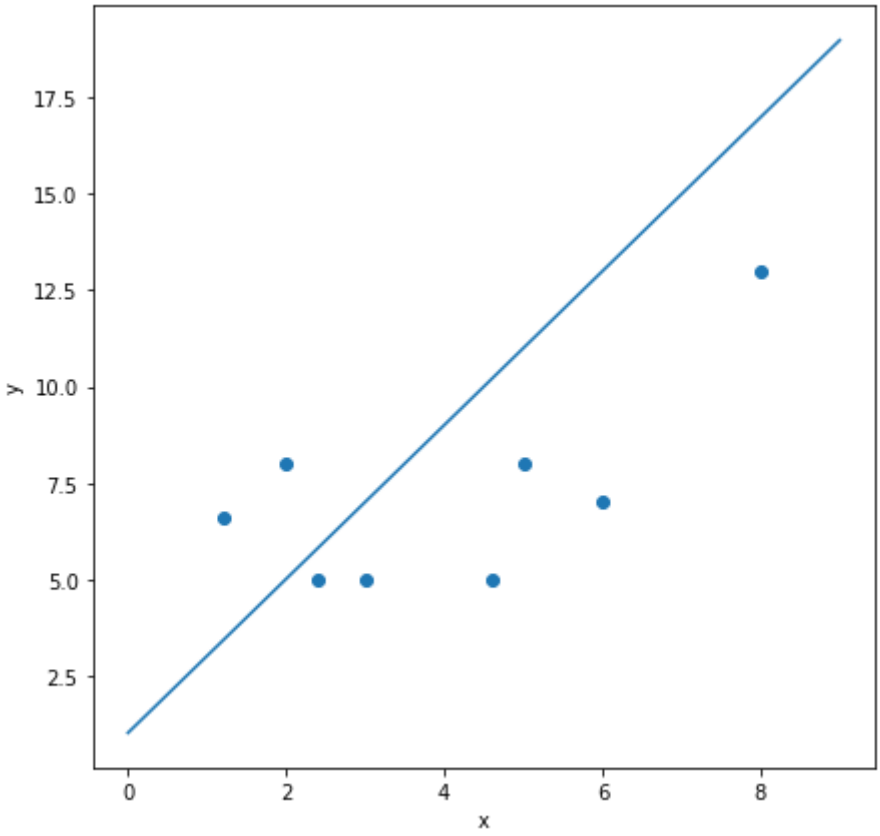
In [4]:

```
import numpy as np
from matplotlib import pyplot as plt

# 임의의 7개 점
z = [5, 2.4, 6, 1.2, 4.6, 2, 8, 3]
w = [8, 5, 7, 6.6, 5, 8, 13, 5]

# 일차 함수 자료 생성, x 구간을 xmin ~ xmax
xmin = 0
xmax = 10
x = np.arange(xmin, xmax)
y = 2 * x + 1

# 한 플롯에 동시 그리기
plt.figure(figsize = (7, 7)) # figure의 size를 7 by 7로 키우기
plt.scatter(z, w)
plt.plot(x, y)
plt.xlabel("x"); plt.ylabel("y")
plt.show()
```



==========

1.4. seaborn.residplot을 이용하여 위 회귀 모형의 잔차 분석을 해보자. (model.fittedvalues, model.resid 를 이용하면 쉽게 그릴 수 있다.) 잔차의 절댓값이 가장 큰 관측값 10개를 뽑아 그래프에 표시해보자. 그리고 잔차도를 통해 위에서 진행한 단순 회귀 모형이 타당한지 간단히 기술하여 보자.

예제 2

wine.csv 데이터는 kaggle에서 다운로드할 수 있는 자료로 레드 와인의 여러 특성을 담은 데이터이다. 우린 이 가운데에서 pH, alcohol, fixedacidity, residualsugar 를 이용한다. 이를 다음과 같이 불러오고 다음 물음들에 답하라.

In [5]:

```
import pandas as pd
wine = pd.read_csv("wine.csv")
```

2.1. wine에서 y = wine["pH"], x = wine["fixedacidity"]로 두고 단순 회귀 분석을 진행해보자. 이 때 F 검정 혹은 t 검정을 이용해 선형 관계가 타당한지 파악해보라. (힌트 : Prob(F-statistic) 혹은 P>|t|를 이용)

2.2 문제 1.4에서 했던 것과 같이 추정된 선형 식과 산점도를 동시에 하나의 플롯에 담아보자. 선형 식에서의 정의역은 np.arange(4, 13)로 두자.

2.3. stats.pearsonr를 이용해 와인의 pH와 알코올 양은 상관관계가 있는지 wine["pH"]와 wine["alcohol"]를 이용해 피어슨 상관계수를 구해보고, 그 크기는 얼마이며, p 값을 통해 어느 정도 타당한지 파악해보자.

2.4 와인이 가지는 여러 성질들을 이용하여 quality를 예측하고자 한다.
$y = fixedacidity, x1 = pH, x2 = alcohol, x3 = residualsugar$ 로 두자. ols 함수를 이용하여 중회귀 분석을 진행해보자. (즉, $y = \beta_0 + \beta_1 x1 + \beta_2 x2 + \beta_3 x3 + \epsilon$) 그리고 F 혹은 t 검정을 이용하여 회귀분석 결과의 타당성을 주장해보자.

2.5. seaborn.pairplot을 이용하여 4 by 4 산점도 행렬을 그려보고, 추가로 data.corr 메소드를 사용하여 (data는 데이터프레임 형식) 4 by 4 피어슨 상관계수 행렬을 생성하여 살펴보자. 상관계수 값을 보고 각 변수간 상관관계가 존재하는지, 결과를 보고 어떤 해석을 할 수 있는지에 대해 간략히 적어보자.

In [6]:

```
data = wine[["fixedacidity", "pH", "citricacid", "residualsugar"]]
```

2.6 문제 2.3과 같이 stats.pearsonr을 이용해 pH와 residual sugar 간 상관관계가 있다고 할 수 있는지 판단하여라. (pearson 상관계수를 두 변수에 관해 다시 구하고, p 값을 이용해 판단해보자.)