선형회귀 예시 (2) - 의료비에 영향을 미치는 요소

#01. 작업 준비

1. 패키지 참조하기

```
import sys
sys.path.append("../../")
import helper

import numpy as np
from pandas import read_excel, DataFrame, melt, merge
from pca import pca
from pandas.api.types import CategoricalDtype
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
from matplotlib import pyplot as plt
import seaborn as sb
from scipy import stats
import statsmodels.api as sm
```

2. 데이터 가져오기

미국 환자의 의료비가 들어 있는 데이터셋으로 `1,338` 개의 관측치가 있다.

변수	의미	기타
age	수익자의 연령	수치형
sex	계약자의 성별	범주형 데이터(female/male)
bmi	미만도. 몸무게를 키의 제곱으로 나눈 값.	수치형 정상범위: 18.5~24.9
children	의료보험이 적용되는 자녀수	수치형 데이터
smoker	흡연 여부	범주형 데이터(yes/no)
region	거주지역	범주형 (북동: northeast, 남동: southeast / 남서: southwest / 북서: northwest)
expense	의료비	수치형 데이터

```
origin = read_excel("https://data.hossam.kr/E04/insurance.xlsx")
origin
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

	age	sex	bmi	children	smoker	region	charges
...	...	...	...	...	...	...	...
1333	50	male	30.970	3	no	northwest	10600.54830
1334	18	female	31.920	0	no	northeast	2205.98080
1335	18	female	36.850	0	no	southeast	1629.83350
1336	21	female	25.800	0	no	southwest	2007.94500
1337	61	female	29.070	0	yes	northwest	29141.36030

1338 rows × 7 columns

#02. 데이터 전처리

1. 데이터 프레임 복사 후 결측치와 데이터 타입 확인

```
edf = origin.copy()
helper.prettyPrint(edf.isna().sum(), title="결측치 개수")
helper.prettyPrint(edf.dtypes, title="데이터 타입")
```

```
+----------+---------------+
|          |   결측치 개수 |
|----------+---------------|
| age      |             0 |
| sex      |             0 |
| bmi      |             0 |
```

```
| children |              0 |
| smoker   |              0 |
| region   |              0 |
| charges  |              0 |
+----------+----------------+
+----------+----------------+
|          | 데이터 타입    |
|----------+----------------|
| age      | int64          |
| sex      | object         |
| bmi      | float64        |
| children | int64          |
| smoker   | object         |
| region   | object         |
| charges  | float64        |
+----------+----------------+
```

2. 범주형 타입 변환

범주형 필드 이름

```
cnames = ["sex", "smoker", "region"]
cnames
```

```
['sex', 'smoker', 'region']
```

범주형 컬럼 타입 변환

```
edf2 = helper.setCategory(edf, fields=cnames, labelling=False)
helper.prettyPrint(edf2.dtypes, title="데이터 타입")
```

```
+----------+---------------+
|          | 데이터 타입   |
|----------+---------------|
| age      | int64         |
| sex      | category      |
| bmi      | float64       |
| children | int64         |
| smoker   | category      |
| region   | category      |
| charges  | float64       |
+----------+---------------+
```

#03. 탐색적 데이터 분석

1. 수치형 변수

1) 기초 통계량 확인

수치형 데이터 타입은 전체적인 통계값을 파악하는 것이 좋다.

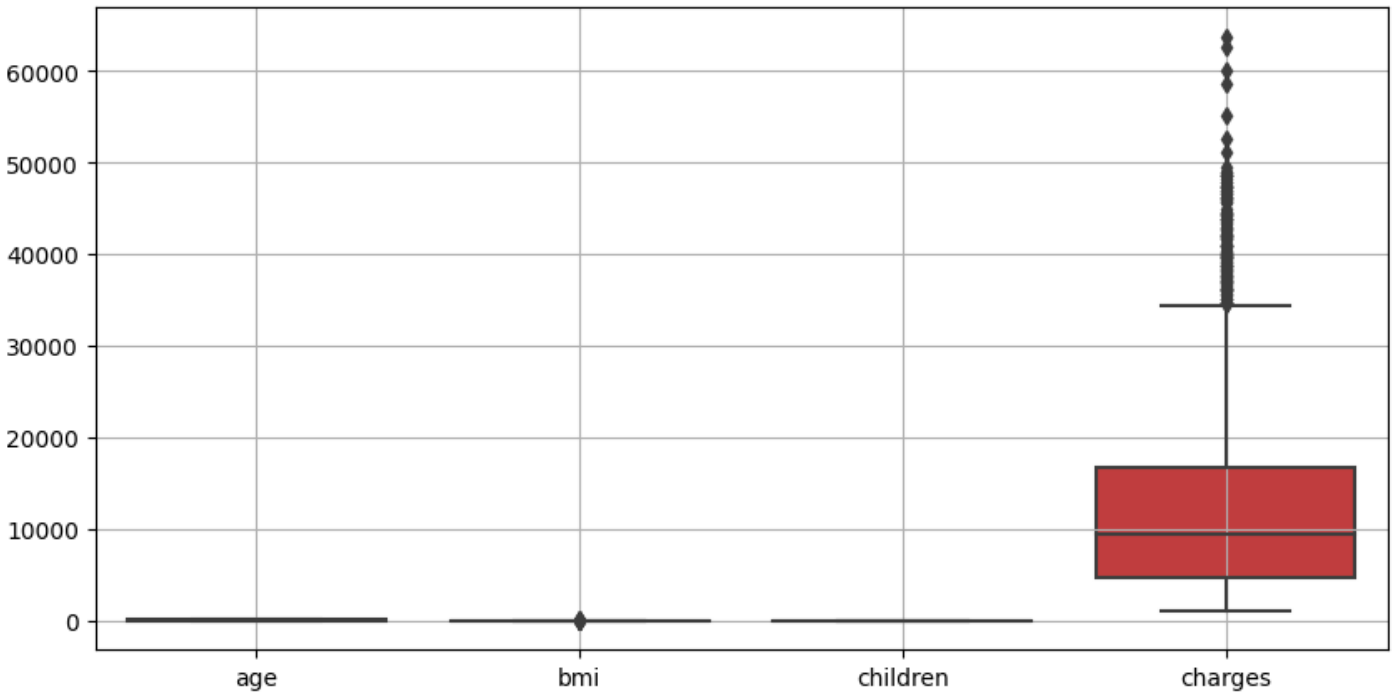
```
desc = edf2.describe()
helper.prettyPrint(desc)
```

	age	bmi	children	charges
count	1338	1338	1338	1338
mean	39.207	30.6634	1.09492	13270.4
std	14.05	6.09819	1.20549	12110
min	18	15.96	0	1121.87
25%	27	26.2963	0	4740.29
50%	39	30.4	1	9382.03
75%	51	34.6938	2	16639.9
max	64	53.13	5	63770.4

> 의료비지출 변수의 통계값을 살펴보면 중앙값이 `$9,382` 이고, 평균이 `$13,270` 인 것을 알 수 있다. 여기서 해당 변수의 평균값이 중앙값보다 크기 때문에 의료비 분포는 오른쪽으로 꼬리가 긴 분포를 지닐 것이다.

2) 전체 상자그림 확인

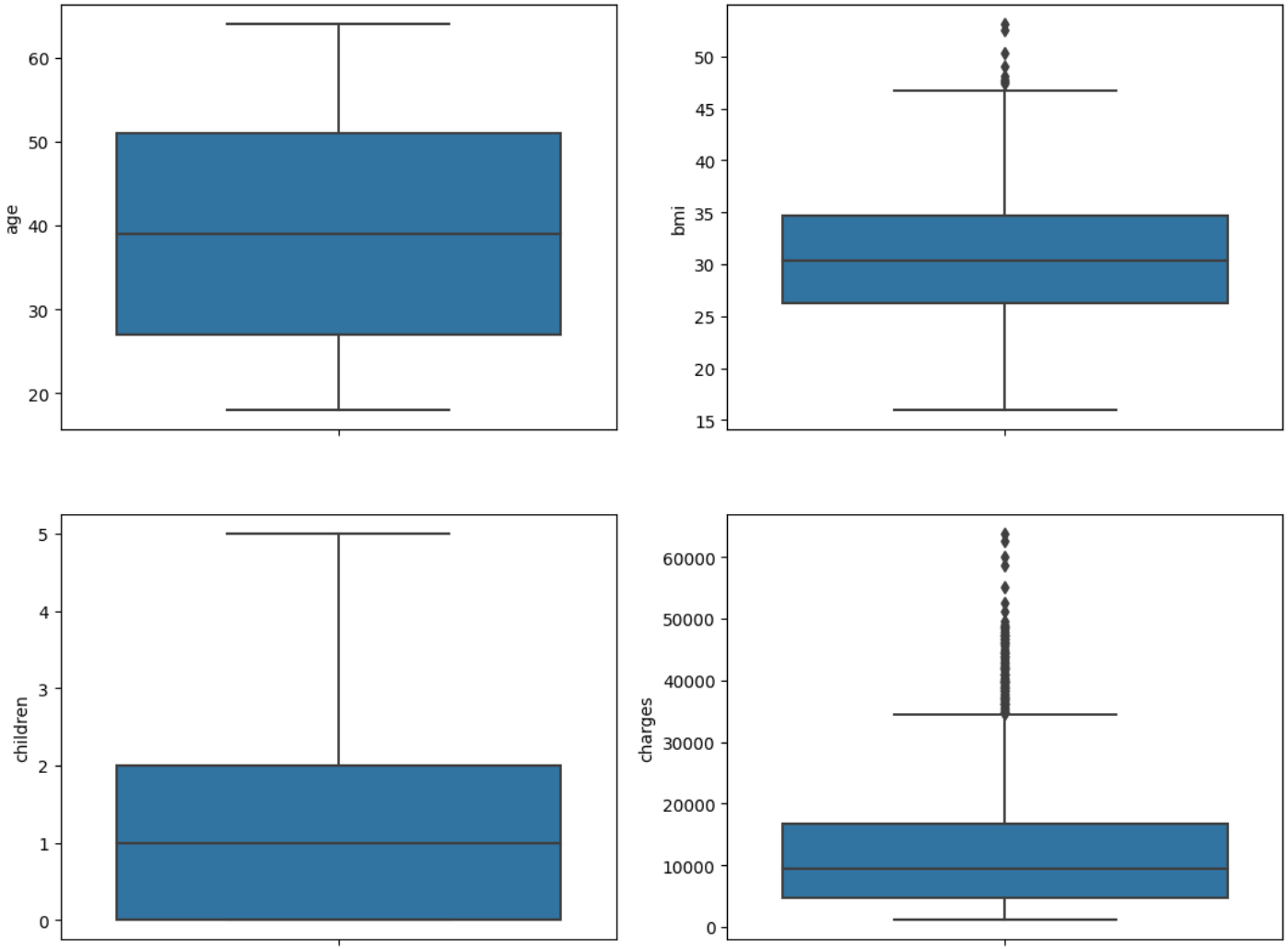
```
plt.figure(figsize=(10, 5))
sb.boxplot(data=edf)
plt.grid()
plt.show()
plt.close()
```



3) 개별 상자그림 확인

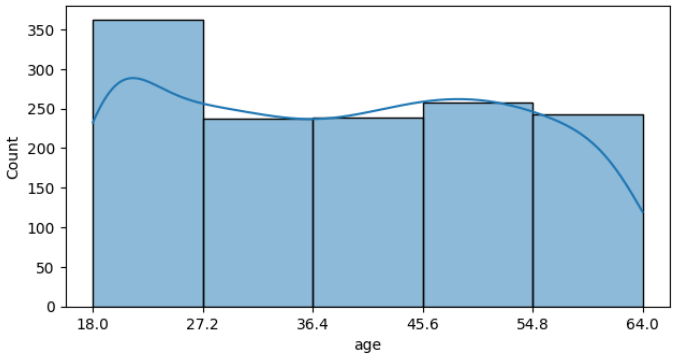
```
fig, ax = plt.subplots(2,2, figsize=(13, 10))
rows = len(ax)
cols = len(ax[0])

for i in range(0, rows):
    for j in range(0, cols):
        idx = i * cols + j
        fieldName = desc.columns[idx]
        field = edf2[fieldName]
        sb.boxplot(edf, y=field, ax=ax[i][j])
```

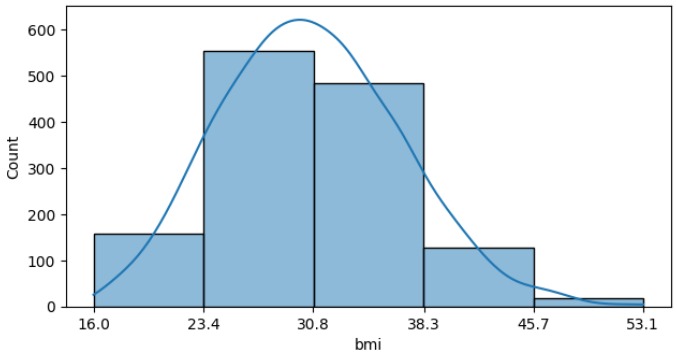
```
        if idx+1 == len(desc.columns):
            break

plt.show()
plt.close()
```

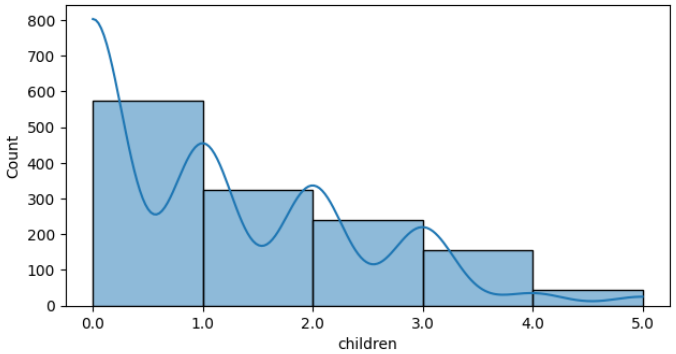


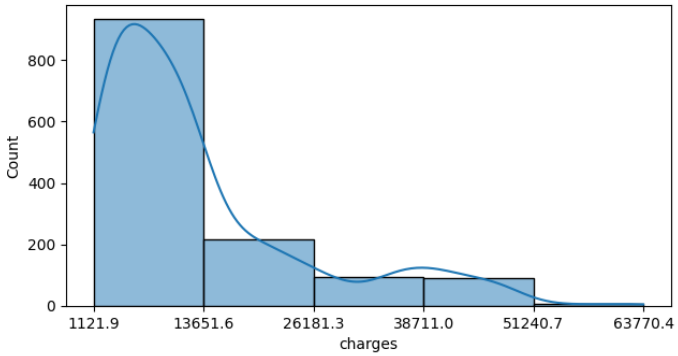
4) 히스토그램 확인

```
fig, ax = plt.subplots(2, 2, figsize=(16, 8))

rows = len(ax)
cols = len(ax[0])

for i in range(rows):
    for j in range(cols):
        idx = i * cols + j
        fieldName = desc.columns[idx]
        field = edf[fieldName]

        hist, bins = np.histogram(field, bins=5)
        bins2 = np.round(bins, 1)

        sb.histplot(data=edf2, x=fieldName, bins=5, kde=True, ax=ax[i][j
        ax[i][j].set_xticks(bins2)
        ax[i][j].set_xticklabels(bins2)

        if idx+1 == len(desc.columns):
            break
plt.show()
plt.close()
```









> 대부분의 사람들은 연간 $15,000 이하의 의료비 지출에 분포되어 있음을 알 수 있다.
>
> 마찬가지로 bmi지수를 살펴보면 과체중이상의 데이터가 절반 이상을 차지하는 것을 알 수 있다.

2. 범주형 변수

1) 종류별로 데이터 수량 확인

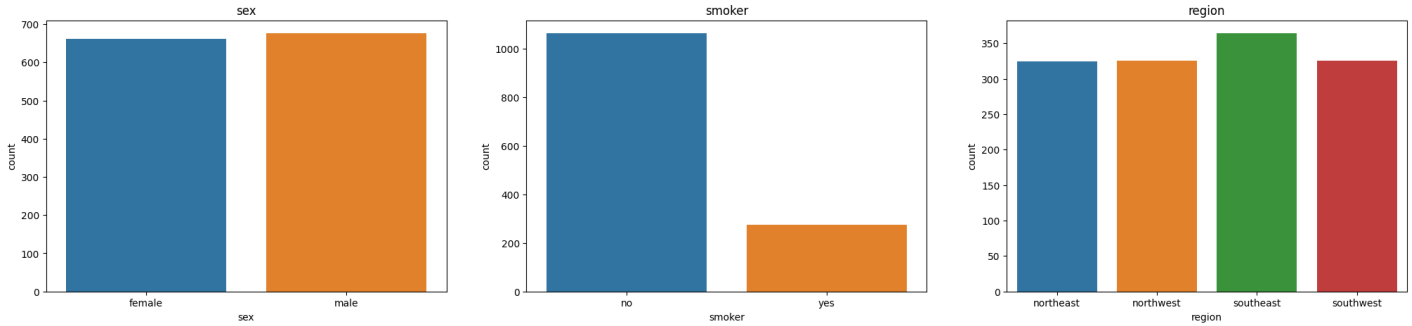
```
for name in cnames:
    helper.prettyPrint(edf2[name].value_counts(), title="count")
```

```
+--------+---------+
| sex    |   count |
|--------+---------|
| male   |     676 |
| female |     662 |
+--------+---------+
+---------+---------+
| smoker  |   count |
|---------+---------|
| no      |    1064 |
| yes     |     274 |
+---------+---------+
+-----------+---------+
| region    |   count |
|-----------+---------|
| southeast |     364 |
| northwest |     325 |
| southwest |     325 |
| northeast |     324 |
+-----------+---------+
```

2) 범주형 데이터의 데이터 분포 시각화

```
fig, ax = plt.subplots(1, len(cnames), figsize=(25, 5))

for i, v in enumerate(cnames):
    vc = DataFrame(edf2[v].value_counts(), columns=['count'])
    #print(vc)
```

```
    sb.barplot(data=vc, x=vc.index, y='count', ax=ax[i])
    ax[i].set_title(v)

plt.show()
plt.close()
```



흡연 여부의 경우 비흡연자가 많이 분포되어 있다.

그 밖에 성별과 지역의 경우 비슷하게 분포되어 있기 때문에 분산분석을 통해 통제요인으로 넣는 것을 고려해 볼 수 있겠다.