ADP_Red\stats\4.regression.py

```python

# 4.Regression
# %% 0. Import Libaries
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

import scipy.stats as stats


# %% 1. Load Dataset
df = pd.read_csv('../ADP_Python/data/insurance.csv')

# Check dataset and set dependent variable(dv) and independent variable(iv)
print(f'{df.info()}')
dv = 'charges'
iv = 'age'


# %% 2. Data Visualization
sns.scatterplot(df, x=iv, y=dv)
plt.tight_layout
plt.show()


# %% 3. Statistical Test (Statsmodels)
import statsmodels.api as sm
import statsmodels.formula.api as smf

lm = smf.ols(
    formula = f'{dv} ~ 1 + {iv}', data=df
).fit()

lm.rsquared
lm.rsquared_adj
lm.params
lm.pvalues

print(lm.summary())
print(f'''
    나이를 사용한 의료비용 예측을 위해 단순 선형 회귀 분석을 시행하였다.
    모형의 p value는 {lm.f_pvalue:.2f}로 모형이 데이터를 잘 설명한다.
    모형의 결정계수(R-squared)는 {lm.rsquared:.3f}, 수정결정계수(Adjusted R-squared)는 {lm.rsquared_adj:.3f}다.

    {iv} 변수의 절편은 {lm.params.iloc[1]:.2f}, p value는 {lm.pvalues.iloc[1]:.2f}로
    통계적으로 유의미하다.
''')


# %% 4. Statistical Test (Scikit-learn SGDRegressor)
from sklearn.linear_model import SGDRegressor

df_iv = np.array(df[iv]).reshape(-1,1)
df_dv = np.array(df[dv]).reshape(-1,1)

```

```python
lm = SGDRegressor(max_iter=1000, random_state=34)
lm.fit(df_iv, df_dv)

print(f'''
    {lm.intercept_}
    {lm.coef_}
''')

# %% 5. Evaluation


```