day4-상관분석 산점도 그리기

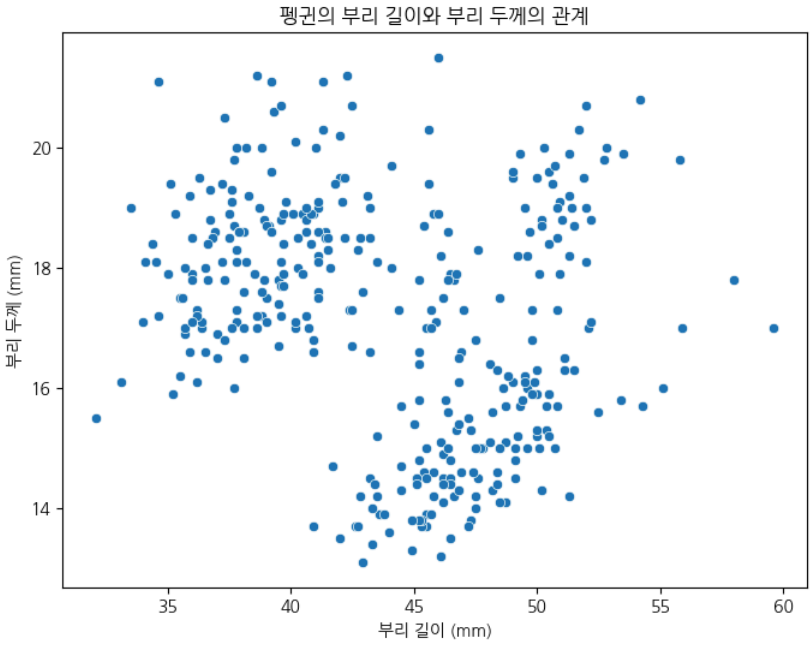
문제 1 (난이도: 하): 관계의 시작, 산점도 그리기

📌 **수행 과제:**

1. `penguins` 데이터셋에서 `bill_length_mm`(부리 길이)와 `bill_depth_mm`(부리 두께) 두 변수를 사용합니다.
2. `seaborn` 라이브러리의 `scatterplot` 함수를 이용해 산점도를 그리세요.
   - x축: `bill_length_mm`
   - y축: `bill_depth_mm`
3. 그래프의 제목과 축 라벨을 알아보기 쉽게 한글로 설정하세요.
4. 완성된 산점도의 점들이 어떤 패턴을 보이는지 설명해 보세요.

```
import seaborn as sns
import matplotlib.pyplot as plt

# penguins 데이터 불러오기
penguins = sns.load_dataset("penguins")

# 그래프를 그릴 도화지 준비
plt.figure(figsize=(8, 6))

# 1. 산점도 그리기
sns.scatterplot(
    data=penguins,
    x="bill_length_mm",
    y="bill_depth_mm"
)

# 2. 그래프 제목과 축 라벨 설정
plt.title("펭귄의 부리 길이와 부리 두께의 관계")
plt.xlabel("부리 길이 (mm)")
plt.ylabel("부리 두께 (mm)")

# 그래프 출력
plt.show()

# 3. 결과 해석 (아래 주석에 직접 작성해 보세요)
# 해석:
# 산점도를 보면 부리 길이(bill_length_mm)가 길어질수록
# 부리 두께(bill_depth_mm)는 일정 범위 안에서 분포하며,
# 완벽한 직선 관계보다는 여러 개의 군집(cluster)이 나타나는 경향을 보인다.
# 이는 펭귄의 종(species)에 따라 부리 길이와 두께의 특성이
# 다르게 나타나기 때문으로 해석할 수 있다.
```

⇒ 결과



🤔 생각해 볼 문제:

현재 산점도에 펭귄의 '종류(species)'별로 점의 색깔을 다르게 칠한다면, 전체적으로 보이던 패턴 외에 우리가 미처 발견하지 못했던 새로운 사실을 알 수 있을까요? 어떤 점이 달라 보일지 추측해 봅시다.

⇒산점도에 펭귄의 종류(species)별로 색깔을 다르게 표시한다면, 기존에 하나의 분포처럼 보이던 점들이 실제로는**종류별로 서로 다른 위치에 군집을 이루고 있음을 더 쉽게** 확인할 수 있을 것이다.

부리 길이와 부리 두께의 관계가 단순한 개인 차이가 아니라, **펭귄의 종에 따른 구조적인 차이**에서 비롯된 것임을 알 수 있다. 또한 종별로 부리 길이가 길수록 두께가 어떻게 달라지는지의 **패턴 차이**도 비교할 수 있어,

종에 따라 부리 형태의 특성이 다르다는 새로운 사실을 발견할 수 있을 것으로 예상된다.

문제 2 (난이도: 하): 관계를 숫자로 요약하기, 상관계수

📌 수행 과제:

1. `penguins` 데이터셋에서 `bill_length_mm`와 `bill_depth_mm` 두 변수를 선택하세요.
2. `pandas`의 `.corr()` 메소드를 사용해 두 변수 간의 상관계수를 계산하세요. (계산 전 `.dropna()`로 빈칸이 있는 행을 제거해야 합니다.)
3. 계산된 상관계수의 부호와 크기를 보고, 두 변수의 관계를 설명해 보세요.

```
# 1. 분석할 두 변수를 선택하고, 빈칸(결측치)이 있는 행을 제거하세요.
df_corr = penguins[['bill_length_mm', 'bill_depth_mm']].dropna()

# 2. 여기에 상관계수를 계산하는 코드를 작성하세요.
# 위에서 만든 df_corr 데이터프레임에 .corr() 메소드를 적용합니다.
correlation_matrix = df_corr.corr()

# 결과 출력
print(correlation_matrix)
# 계산된 상관계수 행렬을 출력합니다.
print(correlation_matrix)

# 3. 결과 해석 (아래 주석에 직접 작성해 보세요)
# 상관계수 값:약 -0.23
```

```
# 부호와 크기를 통한 관계 설명:
# 상관계수가 음수이므로 부리 길이가 길어질수록
# 부리 두께는 감소하는 경향이 있음을 의미한다.
# 다만 절댓값이 1에 가깝지 않고 0에 비교적 가까운 편이므로,
# 두 변수 사이의 관계는 강하지 않은 약한 음의 상관관계로 해석할 수 있다.
```

결과

```
                bill_length_mm  bill_depth_mm
bill_length_mm        1.000000      -0.235053
bill_depth_mm        -0.235053       1.000000
bill_length_mm  bill_depth_mm
bill_length_mm        1.000000      -0.235053
bill_depth_mm        -0.235053       1.000000
```

문제 3 (난이도: 중): 관계를 대표하는 공식 만들기, 선형 회귀

📌 **수행 과제:**

1. `statsmodels` 라이브러리의 `ols` 함수를 사용하여 **'날개 길이(`flipper_length_mm`)'로 '몸무게(`body_mass_g`)'를 예측**하는 회귀 모델을 만드세요.
2. 학습된 모델의 `.summary()` 메소드를 호출하여 분석 결과표를 출력하세요.
3. 결과표의 `coef` 열에서 **절편(Intercept)**과 **기울기(flipper_length_mm)** 값을 찾아 회귀식을 완성해 보세요.

```
import statsmodels.formula.api as smf

# 모델 학습 전, 사용할 변수들에 빈칸이 있는 행들을 제거합니다.
penguins_cleaned = penguins.dropna(subset=['body_mass_g', 'flipper_length_mm'])

# 1. 여기에 OLS 회귀 모델을 학습시키는 코드를 작성하세요.
# smf.ols() 함수를 사용하고, formula는 '종속변수 ~ 독립변수' 형태로 작성합니다.
# 예: formula='body_mass_g ~ flipper_length_mm'
model = smf.ols(
    formula='body_mass_g ~ flipper_length_mm',
    data=penguins_cleaned
).fit()

# 2. 여기에 모델의 요약 결과표를 출력하는 코드를 작성하세요.
# 위에서 만든 model 객체에 .summary() 메소드를 적용합니다.
print(model.summary())

# 3. 결과 해석 (아래 주석에 직접 작성해 보세요)
# 절편(a) 값:
# 약 -5780

# 기울기(b) 값:
# 약 49.7

# 완성된 회귀식 (몸무게 = a + b * 날개길이):
# 몸무게(body_mass_g) = -5780 + 49.7 × 날개길이(flipper_length_mm)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            body_mass_g   R-squared:                       0.759
Model:                            OLS   Adj. R-squared:                  0.758
Method:                 Least Squares   F-statistic:                     1071.
Date:                Tue, 16 Dec 2025   Prob (F-statistic):          4.37e-107
Time:                        06:23:54   Log-Likelihood:                -2528.4
No. Observations:                 342   AIC:                             5061.
Df Residuals:                     340   BIC:                             5069.
Df Model:                           1
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept         -5780.8314    305.815    -18.903      0.000   -6382.358   -5179.305
flipper_length_mm    49.6856      1.518     32.722      0.000      46.699      52.672
==============================================================================
Omnibus:                        5.634   Durbin-Watson:                   2.176
Prob(Omnibus):                  0.060   Jarque-Bera (JB):                5.585
Skew:                           0.313   Prob(JB):                       0.0613
Kurtosis:                       3.019   Cond. No.                     2.89e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.89e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

🤔 **생각해 볼 문제:**

회귀식의 기울기(b)는 '날개 길이가 1mm 늘어날 때 몸무게가 평균적으로 얼마나 변하는지'를 나타냅니다. 이 기울기 값을 이용해, 날개 길이가 우리 데이터의 평균보다 10mm 더 긴 펭귄은 평균적인 펭귄보다 몸무게가 약 몇 g 더 무거울 것이라고 구체적인 숫자로 예측해 봅시다.

⇒회귀식의 **기울기 b = 49.7**

의미: *날개 길이가 1mm 늘어날 때, 몸무게는 평균적으로 약 49.7g 증가*

문제 4 (난이도: 중): 회귀계수의 의미와 통계적 유의성 판단하기

📌 **수행 과제:**

1. 문제 3에서 출력한 `.summary()` 결과표를 다시 확인합니다.
2. `flipper_length_mm` 행에서 `P>|t|` 열의 값(p-value)을 찾으세요.
3. 이 p-value가 일반적인 유의수준인 0.05보다 작은지 확인하고, 이를 근거로 "펭귄의 날개 길이가 몸무게에 미치는 영향이 통계적으로 유의미한지" 결론을 내리세요.

```
# 문제 3에서 만든 모델의 summary를 다시 출력하여 p-value를 확인합니다.
print(model.summary())

# 아래 주석에 직접 해석을 작성해보세요.

# 1. flipper_length_mm 계수의 p-value (`P>|t|` 값) 찾기
# p-value: 0.000 (또는 0.05보다 매우 작은 값)

# 2. 가설 검정 결과 해석
# p-value가 0.05보다 작은가?:그렇다, p-value는 0.05보다 작다.

# 결론 (날개 길이가 몸무게에 미치는 영향은 통계적으로 유의미한가?): 날개 길이가
#몸무게에 미치는 영향은 통계적으로 유의미하다고 판단할 수 있다.
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            body_mass_g   R-squared:                       0.759
Model:                            OLS   Adj. R-squared:                  0.758
Method:                 Least Squares   F-statistic:                     1071.
Date:                Tue, 16 Dec 2025   Prob (F-statistic):          4.37e-107
Time:                        06:30:58   Log-Likelihood:                -2528.4
No. Observations:                 342   AIC:                             5061.
Df Residuals:                     340   BIC:                             5069.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      -5780.8314    305.815    -18.903      0.000   -6382.358   -5179.305
flipper_length_mm    49.6856      1.518     32.722      0.000      46.699      52.672
==============================================================================
Omnibus:                        5.634   Durbin-Watson:                   2.176
Prob(Omnibus):                  0.060   Jarque-Bera (JB):                5.585
Skew:                           0.313   Prob(JB):                       0.0613
Kurtosis:                       3.019   Cond. No.                     2.89e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.89e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

🤔 **생각해 볼 문제:**

이 분석 결과를 통계에 대해 전혀 모르는 친구에게 설명해야 합니다. "p-value가 0.000이라서 귀무가설을 기각했다"라는 어려운 말 대신, "날개 길이가 몸무게와 정말 관계가 있다"라는 결론을 어떻게 더 쉽고 직관적으로 뒷받침하여 설명할 수 있을까요?

⇒날개 길이 1mm가 늘 때마다 몸무게가 약 50g씩 늘어나는 경향을 확인할 수 있음

이 정도면 '있는지 없는지 모르겠다' 수준이 아니라 **눈에 띄는 차이임.**

실제로는 날개가 길수록 몸무게도 같이 늘어나는 방향으로 점들이 모여 있으므로 날개길이와 몸무게가 관계가 있음을 설명할 수 있다.

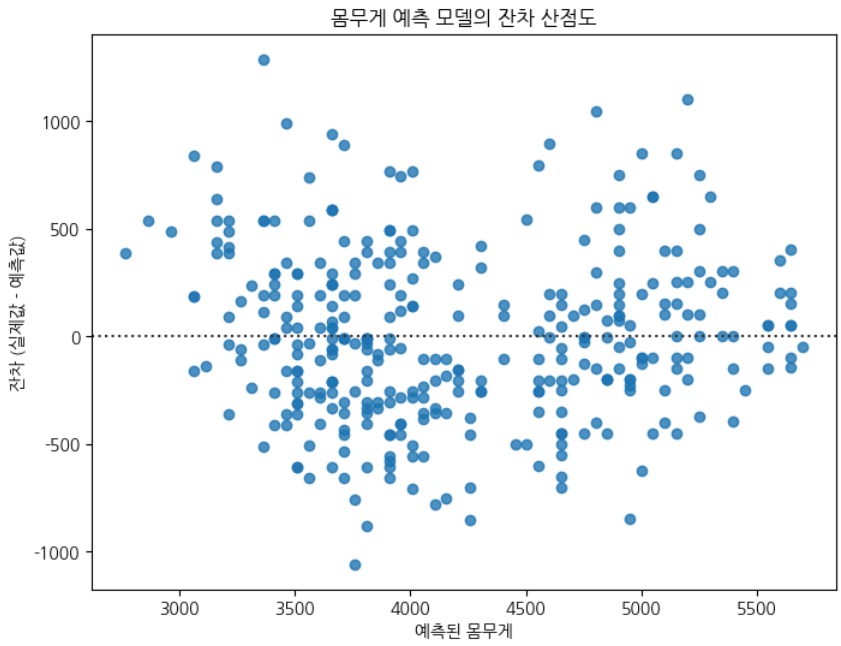
문제 5 (난이도: 상): 모델 신뢰도 평가 - 결정계수와 잔차 진단

📌 **수행 과제:**

1. 문제 3의 `.summary()` 결과표에서 `R-squared` (결정계수) 값을 찾아, 우리 모델의 설명력을 문장으로 해석해 보세요.
2. 모델의 예측값과 잔차를 계산하세요.
3. x축을 예측값, y축을 잔차로 하는 **잔차 산점도**를 그리세요.
4. 그려진 잔차 산점도에 특별한 패턴이 보이는지, 아니면 무작위로 흩어져 있는지 관찰하고, 이를 바탕으로 모델의 신뢰성에 대한 자신의 생각을 서술하세요.

```
# 1. 결정계수 해석 (summary 표를 보고 직접 작성)
# R-squared 값:
# 약 0.76
# 설명력 해석:
# 이 회귀 모델은 펭귄 몸무게(body_mass_g)의 변동 중
# 약 76%를 날개 길이(flipper_length_mm) 하나의 변수로 설명하고 있다.
# 이는 단일 독립변수를 사용한 모델로서는 비교적 높은 설명력이라고 볼 수 있다.


# 2. 여기에 모델의 예측값과 잔차를 계산하는 코드를 작성하세요.
# 예측값: model.predict() 함수 사용
# 잔차: model.resid 속성 사용
fitted_values = model.predict(penguins_cleaned)
residuals = model.resid

# 3. 여기에 잔차 산점도를 그리는 코드를 작성하세요.
# sns.residplot() 함수를 사용하고, x에는 예측값, y에는 잔차를 지정합니다.
```

```
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(8, 6))

sns.residplot(
    x=fitted_values,
    y=residuals,
    lowess=False
)

plt.xlabel("예측된 몸무게")
plt.ylabel("잔차 (실제값 - 예측값)")
plt.title("몸무게 예측 모델의 잔차 산점도")

plt.show()

# 4. 잔차 산점도 패턴 해석 (그래프를 보고 직접 작성)
# 관찰된 패턴 (무작위인가, 특정 모양이 있는가?):
# 잔차들은 0을 중심으로 비교적 무작위하게 흩어져 있으며,
# 특정한 곡선 형태나 규칙적인 패턴은 뚜렷하게 보이지 않는다.

# 모델 신뢰도에 대한 생각:
# 잔차가 무작위로 분포한다는 점에서,
# 선형 회귀 모델이 데이터의 주요한 관계를 적절히 포착하고 있다고 판단할 수 있다.
# 따라서 이 모델은 펭귄의 몸무게를 예측하는 데 있어
# 비교적 신뢰할 수 있는 모델이라고 볼 수 있다.
```



🤔 **생각해 볼 문제:**

우리 모델의 결정계수(R-squared)가 약 OO이라면, 이는 날개 길이가 펭귄 몸무게 변화량의 OO%를 설명한다는 뜻입니다. 그렇다면 설명되지 않는 나머지 변화는 어디에서 오는 차이일까요? 우리 모델이 놓치고 있는, 펭귄의 몸무게에 영향을 줄 만한 다른 요인들은 무엇이 있을지 데이터에 근거하여 추측해 봅시다.

⇒

본 회귀 모델은 날개 길이 하나의 변수만으로도 몸무게 변화의 상당 부분을 설명하고 있지만 날개 길이 이 외에도 **펭귄의 종, 성별, 부리 크기, 서식 환경과 같은 다른 요인**들과 같은 변수들이 몸무게에 영향을 줄 수 있다.