

05

다중 선형 회귀분석 실습

[다중선형회귀분석 실습 - 전처리]

집값 데이터 예측을 해보겠습니다.

00:00

다. 모형 내의 회귀계수가 유의한가?

- 단변량 회귀분석에서 회귀계수의 유의성 검토와 마찬가지로 회귀계수에 대한 t통계량의 p-value값이 0.05보다 작으면 해당 회귀 계수가 통계적으로 유의하다고 볼 수 있다.
- 단, 다중회귀분석을 할 때는 모든 회귀계수가 유의한지를 검증한 후 해당 회귀식을 해석해야 함

마. 모형이 데이터를 잘 적합하고 있는가?

- 모형의 잔차와 종속변수에 대한 산점도를 그리고, 회귀진단을 수행하여 판단한다.

5.3.2.1 3. 더미변수

가. 범주형 변수 변환

- 회귀분석은 연속형 변수를 다루는 기법이므로 범주형 데이터의 경우 형태를 변환해주어야 회귀분석을 수행할 수 있다.
- 더미변수란 0 or 1값만 가지며 어떤 특징에 해당 하는지의 여부를 표현하는 변수이다.

[예제]

kc_house_data 데이터에서 price를 종속변수로 설정하고, date와 id를 제거한 15개의 컬럼을 독립변수로 설정하여 다중선형 회귀분석을 실시한 후, 추정된 회귀모형에 대해 해석하라

```
In [27]: 1 import pandas as pd
2 import numpy as np
3 house = pd.read_csv('../data/kc_house_data.csv')
4 house.head()
```

[모델링]

03:07

```
In [8]: 1 ols_str[:3]
```

```
Out[8]: 'price ~ bedrooms + bathrooms + sqft_living + sqft_lot + floors + view + condition + grade + sqft_above + sqft_basement + yr_built + yr_renovated + sqft_living15 + sqft_lot15 + waterfront_river_view + waterfront_standar d'
```

```
In [9]: 1 ols_str = ols_str[:3]
```

```
In [37]: 1 import numpy as np
2 import statsmodels.api as sm
3 import statsmodels.formula.api as smf
4
5
6 model = smf.ols(formula = ols_str, data = house)
7 result = model.fit()
8 result.summary()
9
10
11 ## 범주형 변수가 추가되면 오류가 남
12 ## 더미변수로 변환 필요
```

condition_2	1.119e+06	2.76e+04	40.530	0.000	1.06e+06	1.17e+06
condition_3	1.116e+06	2.52e+04	44.223	0.000	1.07e+06	1.17e+06
condition_4	1.129e+06	2.44e+04	46.184	0.000	1.08e+06	1.18e+06
condition_5	1.167e+06	2.39e+04	48.817	0.000	1.12e+06	1.21e+06

Omnibus: 16350.766 Durbin-Watson: 1.979
Prob(Omnibus): 0.000 Jarque-Bera (JB): 1202226.985