**비즈니스 데이터마이닝 2020년 2학기 중간고사**

1. 분석에 필요한 모든 변수명은 임의로 만들어 사용하십시오.
2. 정답을 작성한 중간고사 문제지(doc 또는 hwp)파일과 코드(html)파일을 함께 제출하시기 바랍니다.
3. 코드파일에서 주석처리가 모두 되어 있는 경우, 최대 2 points 가점됩니다.
4. 시험종료(11:30) 전까지 제출파일(2개)을 모두 업로드하시기 바랍니다.

* 시험이 종료되면 제출할 수 없습니다. -

**문제. 보스톤 지역의 주택가격(MEDV)**에 미치는 영향요인을 확인하기 위하여 다중선형회귀를 가정하고 다음 분석을 진행하시오.

(‘bd01\_exam.csv’ , 변수에 대한 설명은 p.3 참고) – 총 20 points

* + 1. NAN값은 평균값으로 대체하시오. 이때 사용한 코드를 적으시오.(3 points)

답:

df['NOX'] = df['NOX'].fillna(df['NOX'].mean())

df['DIS'] = df['DIS'].fillna(df['DIS']).mean()

df['TAX'] = df['TAX'].fillna(df['TAX'].mean())

# nan 값 평균값으로 대체함.

* + 1. 데이터 셋에 대해 상관분석을 진행하고 변수명 ‘MEDV’와 가장 상관관계가 높은 변수 3개를 찾아 적으시오.(4 points)

답:

df.corr()

import seaborn as sns

cat = df.corr()

cat.index

cat.iloc[13].sort\_values(ascending=False)

* **# medv와 상관관계가 가장 높은 변수 세 개는 LSTAT, RM, PTRATIO 이다.**
  + 1. 2)번의 결과를 바탕으로 상관관계가 높은 3개의 변수를 독립변수로 사용하고 MEDV를 종속변수로 하여 다중선형회귀분석을 진행하고 분석결과에 대한 회귀식을 작성하시오.(5 points)

답:

x = df[['LSTAT', 'RM', 'PTRATIO']]

y = df['MEDV']

# 상수항 추가하기

import statsmodels.api as sm

x1 = sm.add\_constant(x, has\_constant='add')

x1.head()

multi\_model = sm.OLS(y, x1).fit()

multi\_model.summary()

# 위와 같은 코드로 다중선형회귀분석 진행하였습니다.

**# 분석결과에 대한 회귀식 : MEDV = 18.5671 + -0.5718\*LSTAT + 4.5154\*RM +**

**-0.9307\*PTRATIO**

* + 1. 3)의 분석결과를 해석하고 주택가격**(MEDV)**에 가장 큰 영향을 미치는 변수는 무엇인지 적으시오. (5 points)

답:

# 분석결과 해석 :

# 이 분석은 67.7% 설명력을 갖는다.

# F분포 p 값이 유의수준 0.05보다 작기 때문에, 이 값은 통계적으로 '타당하다'

# coef값들도 p 값이 유의수준 0.05보다 작기 때문에 이 값들은 통계적으로 '타당하다'

# LSTAT = -0.5718, RM = 4.5154, PTRATIO= -0.9307 만큼 각각 MEDV에 영향 미친다.

# LSTAT가 1단위 증가할때 MEDV는 -0.5718만큼 감소한다.

# RM이 1단위 증가할 때, MEDV는 4.5145만큼 증가한다.

# PTRATIO 1단위 증가할 때 MEDV는 -0.9307만큼 감소한다.

**# 따라서 MEDV, 주택 가격에 가장 큰 영향 미치는 변수는 RM이다.**

* + 1. **4)번 결과를 토대로 가장 큰 영향을 미친 변수를 독립변수로 놓고 주택가격(MEDV)를 종속변수로 하여 회귀선 그래프를 그리시오.(쥬피터 노트북에 실행한 상태로 제출하고, 문제지에는 따로 작성하지 않습니다. 라이브러리는 어떠한 것을 사용해도 좋습니다. )(3 points)**

--------- 수고하셨습니다. ----------

<bd01\_exam.csv데이터 파일 변수 설명>

Boston house prices

(https://archive.ics.uci.edu/ml/machine-learning-databases/housing/)

**Data Set Characteristics:**

**Number of Instances**

506

**Number of Attributes**

13 numeric/categorical predictive.

MEDV is usually the target variable.

**Attribute Information (in order)**

|  |  |
| --- | --- |
| 변수명 | 설명 |
| CRIM | 자치시(town)별 1인당 범죄율 |
| ZN | 25,000 평방피트를 초과하는 거주지역 비율 |
| INDUS | 비소매상업지역이 점유하고 있는 토지 비율 |
| CHAS | 찰스강에 대한 더미변수(강의 경계에 위치한 경우 1, 아니면 0) |
| NOX | 농축 일산화 질소(parts per 10 million) |
| RM | 주택 1가구당 평균 방의 개수 |
| AGE | 1940년 이전에 건축된 소유주택의 비율 |
| DIS | 5개의 보스턴 직업센터까지의 접근성 지수 |
| RAD | 방사형 도로까지의 접근성 지수 |
| TAX | 재산세율 |
| PTRATIO | 자치시(town)별 학생/교사 비율 |
| B | 자치시(town)별 흑인비율 |
| LSTAT | 모 집단의 하위계층 비율 |
| MEDV | 본인 소유의 주책가격(중앙값, 단위: $1,000)(종속변수) |