応用データ解析　中間課題　　　　　　　12月9日

担当教員：　米倉

PSIコース

学籍番号03-220977

氏名　陳品丞

本課題では添付のデータを使用してもよいし、各自で用意したデータを使用してもよい。

ただし各自のデータを使用する場合は以下の条件を満たすこと。

・標本数が100以上であること。

・架空のデータでなく、実際の調査をもとに作成されたデータであること。（自分で調査していなくてもよい。）

レポートの形式は自由とする。（Wordに結果をまとめても良いし、Jupyter notebookなどで計算結果と一緒に考察を書いても良い。）

添付のデータはボストン市の住宅価格とそれに関係するデータである。変数の意味は次のとおりである。（出典：カーネギーメロン大　http://lib.stat.cmu.edu/datasets/boston）

CRIM per capita crime rate by town

ZN proportion of residential land zoned for lots over 25,000 sq.ft.

INDUS proportion of non-retail business acres per town

CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)

NOX nitric oxides concentration (parts per 10 million)

RM average number of rooms per dwelling

AGE proportion of owner-occupied units built prior to 1940

DIS weighted distances to five Boston employment centres

RAD index of accessibility to radial highways

TAX full-value property-tax rate per $10,000

PTRATIO pupil-teacher ratio by town

B 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town

LSTAT % lower status of the population

MEDV Median value of owner-occupied homes in $1000's

1. MEDV（または各自のデータ）について、以下の量を求めよ。また、これらの用語を必要なだけ使ってMEDVがどのような分布をしているか簡単に説明せよ。

小数第２位までで

・期待値：22.53・中央値：21.20・標本分散：84.42・不偏分散：84.59

・歪度：1.10・尖度：1.47

歪度が正であることから正規分布と比べて左に偏っている、尖度が正であることから正規分布と比べて山なりに分布している。中央値が期待値に比べ小さいことからも、期待値と中央値が同じ値となる正規分布に比べて、分布が左に偏っていると考えられる。

1. MEDV（または各自のデータ）と相関が強い変数と弱い変数を調べ、MEDVとの因果関係の有無を推測して簡単に述べよ。（例：税金を多く収める住民が住む地区は所得が高い傾向にあるため住宅価格が高い傾向にある。）また、他にデータについて気づいた点があれば述べよ

MEDVとの相関係数

CRIM：-0.388305

ZN：0.360445

INDUS：-0.483725

CHAS：0.175260

NOX ：-0.427321

RM：0.695360

AGE：-0.376955

DIS：0.249929

RAD：-0.381626

TAX：-0.468536

PTRATIO：-0.507787

B：0.333461

LSTAT：-0.737663

相関が強い変数

RM→1住戸あたりの平均部屋数が多い地区は部屋数が多い住宅の方が価格は高くなる傾向にあるため、住宅価格が高い傾向にある。

LSTAT→人口当たり地位が低い率が高いほど、低収入の人が多いため、住宅価格が低い傾向にある。

CHASはダミー変数なので今回、相関係数は無視する

相関が弱い変数

DIS→5つのボストンの雇用センターまでの加重距離は住宅価格とはそれほど関係がない。一番近いの雇用センターとの距離は相関が強い可能性があるが、これは5つの雇用センターとの加重距離であるため相関が弱いと考えられる。

NOXとINDUSの相関係数も0.763651と相関が強かった。これは町当たりの非小売業エーカーの割合が高い地域では工場の数が多くなり、一酸化窒素濃度も高くなるからであると考えることができる。

以上