2024/12/03

回帰分析を用いた

ボストンの犯罪率分析

データ解析第二 | 中央大学理工学部ビジネスデータサイエンス学科14班 | データセット取得先：Kaggle

2024/12/03

目次

背景と目的

住環境から犯罪率を予測し、
犯罪発生率を下げる方法を模索したい。

現状

・アフリカ南部やアメリカなどで犯罪が多いイメージ有
・クルド人による犯罪行為の問題視 (in Japan)

目的

・犯罪行為発生には犯行動機だけでなく住環境も関係するのでは？？
・犯罪の起こりやすい住環境の特徴把握->効率的な見回り(犯罪抑止)

線形回帰モデルとは

線形回帰モデル式とその実装

$$y_i = \alpha + \sum_{j=1}^{p} \beta_j x_{j,i} + \varepsilon_j, \quad \varepsilon_j \sim N(0, \sigma^2)$$

切片　　**偏回帰係数** **...j番目の説明変数の係数**

```r
#Rでの線形回帰モデルの実装
result = lm(被説明変数 ~ 説明変数1, 説明変数2, 説明変数3+... ,data= データ名)
summary(result)
```

線形回帰モデルとは

線形回帰モデルのメリットデメリット

「メリット」

・解釈が容易

各偏回帰係数が目的変数への影響度を示す

・モデルがシンプル

過学習しにくく、変数の影響を理解しやすい

「デメリット」

・多重共線性の影響

説明変数間に強い相関があると信頼性低下

・複雑な相互関係は示せない

非線形関係,高次元の関係は示せない。

利用データの説明

kaggle

Kaggle

ボストンの住宅価格データ

Boston House Prices

Advanced Regression Techniques

注

本来は、住宅価格を予測するコンペのデータであるが、
今回は犯罪発生率を予測するものとして扱った。

Data

CRIM	町別人口一人当たりの犯罪発生率	DIS	5つの雇用センターまでの過重距離
ZN	住宅用地の割合	RAD	高速道路までのアクセス性指数
INDUS	非小売業用地の割合	TAX	固定資産税率
CHAS	チャールズ川に接するかのダミー変数	PTRATIO	町ごとの生徒と教師の比率
NOX	一酸化窒素濃度	B	黒人の割合
RM	平均部屋数	LSTAT	地位の低い人口の割合
AGE	老朽化の進んでいる持ち家の割合	MEDV	居住住宅価格の中央値

Boston House Prices-Advanced Regression Techniques, Kaggle, 2021,
(https://www.kaggle.com/datasets/fedesoriano/the-boston-houseprice-data?resource=download)

分析結果1

全ての説明変数を用いた場合

```
Call:
lm(formula = CRIM ~ ZN + INDUS + CHAS + NOX + RM + AGE + DIS +
    RAD + TAX + PTRATIO + B + LSTAT + MEDV, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-9.924 -2.120 -0.353  1.019 75.051

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.033228   7.234903   2.354 0.018949 *
ZN            0.044855   0.018734   2.394 0.017025 *
INDUS        -0.063855   0.083407  -0.766 0.444294
CHAS         -0.749134   1.180147  -0.635 0.525867
NOX         -10.313535   5.275536  -1.955 0.051152 .
RM            0.430131   0.612830   0.702 0.483089
AGE           0.001452   0.017925   0.081 0.935488
DIS          -0.987176   0.281817  -3.503 0.000502 ***
RAD           0.588209   0.088049   6.680 6.46e-11 ***
TAX          -0.003780   0.005156  -0.733 0.463793
PTRATIO      -0.271081   0.186450  -1.454 0.146611
B            -0.007538   0.003673  -2.052 0.040702 *
LSTAT         0.126211   0.075725   1.667 0.096208 .
MEDV         -0.198887   0.060516  -3.287 0.001087 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.439 on 492 degrees of freedom
Multiple R-squared:  0.454,     Adjusted R-squared:  0.4396
F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16

> vif(res)
      ZN    INDUS     CHAS      NOX       RM      AGE      DIS      RAD      TAX  PTRATIO        B    LSTAT     MEDV
2.325094 3.987753 1.094326 4.551563 2.258113 3.100801 4.289041 7.158834 9.195495 1.984489 1.369741 3.561476 3.772856
```

有意と断定できない因子が多数存在する。

VIFが10を超える変数はなかった。

- 説明変数の削減(有意と言えないもの)
- 変数同士の平均,分散に差が大きい->標準化
- 黒人割合->二乗項
- 非小売業用地割合と低地位割合->交互作用項

分析結果2

変更後

```
Call:
lm(formula = CRIM_ ~ DIS_ + RAD_ + I(B_^2) + MEDV_ + INDUS_LSTAT_,
    data = data_csv)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3478 -0.1784 -0.0162  0.0925  8.8882

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.11935    0.04130  -2.890 0.004025 **
DIS_         -0.07773    0.03834  -2.028 0.043141 *
RAD_          0.47659    0.04197  11.356  < 2e-16 ***
I(B_^2)       0.03347    0.01230   2.721 0.006731 **
MEDV_        -0.13468    0.03679  -3.661 0.000278 ***
INDUS_LSTAT_  0.14263    0.03742   3.812 0.000155 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7461 on 500 degrees of freedom
Multiple R-squared:  0.4488,    Adjusted R-squared:  0.4433
F-statistic: 81.43 on 5 and 500 DF,  p-value: < 2.2e-16

> vif(result4)
        DIS_         RAD_      I(B_^2)        MEDV_ INDUS_LSTAT_
    1.333171     1.597708     1.252071     1.227875     1.045049
```

- 全ての説明変数が有意と言えた(水準5%)
- 雇用センターまでの距離が負の影響
- 高速道路へのアクセス性指数が正の影響
- 黒人の割合の二乗項が正の影響
- 平均居住住宅価格が負の影響
- 非小売業用地と低地位割合に正の影響

- VIFが全て小さい値となっているため
  多重共線性が否定できる。

考察

- 雇用センターまでの距離が負の影響
  - ➔ 犯罪発生率の高い地域に雇用センターが設置してある。
- 高速道路へのアクセス性指数が正の影響
  - ➔ アクセスがいい場所=物流,交通量が多いから相対的に犯罪発生件数増加する。
- 黒人割合に二乗項が正の影響
  - ➔ 黒人割合の増加により犯罪発生率の増加が確認できた。
    ※因果関係は示せないため、黒人の低賃金労働等根本的特徴を調査する必要がある。
- 平均居住住宅価格が負の影響
  - ➔ 高級住宅地では犯罪率が低い
- 非小売業用地と低地位割合の交互作用項に正の影響
  - ➔ 工場や倉庫などによる騒音や低賃金労働等により環境が荒れやすいのではないか
- 線形回帰による分析により住環境の犯罪発生率への影響を
  定量的に比較し分析できた

実装コード

```r
library(car)
data_csv=read.csv("Boston.csv")
data_csv
#標準化した項について考えていく
#標準化したシリーズを変数名_と表記する。
#標準化済み単純説明変数
CRIM_ = (data_csv$CRIM -mean(data_csv$CRIM))/sd(data_csv$CRIM)
ZN_ = (data_csv$ZN -mean(data_csv$ZN))/sd(data_csv$ZN)
INDUS_ = (data_csv$INDUS -mean(data_csv$INDUS))/sd(data_csv$INDUS)
CHAS_ = (data_csv$CHAS -mean(data_csv$CHAS))/sd(data_csv$CHAS)
NOX_ = (data_csv$NOX -mean(data_csv$NOX))/sd(data_csv$NOX)
RM_ = (data_csv$RM -mean(data_csv$RM))/sd(data_csv$RM)
AGE_ = (data_csv$AGE -mean(data_csv$AGE))/sd(data_csv$AGE)
DIS_ = (data_csv$DIS -mean(data_csv$DIS))/sd(data_csv$DIS)
RAD_ = (data_csv$RAD -mean(data_csv$RAD))/sd(data_csv$RAD)
TAX_ = (data_csv$TAX -mean(data_csv$TAX))/sd(data_csv$TAX)
PTRATIO_ =(data_csv$PTRATIO -mean(data_csv$PTRATIO))/sd(data_csv$PTRATIO)
B_ = (data_csv$B -mean(data_csv$B))/sd(data_csv$B)
LSTAT_ = (data_csv$LSTAT -mean(data_csv$LSTAT))/sd(data_csv$LSTAT)
MEDV_ = (data_csv$MEDV -mean(data_csv$MEDV))/sd(data_csv$MEDV)
```

```r
#交互作用項
INDUS_LSTAT_ = INDUS_*LSTAT_
#非線形要素
B_2 = |(B_^2)
#線形回帰
result1 = lm(CRIM_ ~ ZN_+INDUS_+CHAS_+NOX_+RM_+AGE_+DIS_+RAD_+TAX_+PTRATIO_+B_+LSTAT_+MEDV_, data=data)
result2 = lm(CRIM_ ~ ZN_+DIS_+RAD_+B_+LSTAT_+MEDV_, data=data_csv)
#交互作用を入れて考える。
result3 = lm(CRIM_ ~ ZN_+DIS_+RAD_+B_+MEDV_+ INDUS_LSTAT_,
             data=data_csv)
#非線形要素として黒人の割合を考慮する。
result4 = lm(CRIM_ ~ DIS_+RAD_+I(B_^2)+MEDV_+ INDUS_LSTAT_,
             data=data_csv)
#標準化しない場合でも実施する。
INDUS_LSTAT = (data_csv$INDUS)*(data_csv$LSTAT)
result5 = lm(CRIM ~ ZN+DIS+RAD+I(B^2)+MEDV+ INDUS_LSTAT,
             data=data_csv)
#標準化しないでモデリングすると有意でない項が出てしまうため、result4の標準化したモデルが優れているのでは、、
#TAX_RADを使う
TAX_RAD = (data_csv$TAX)*(data_csv$RAD)
result6 = lm(CRIM ~ ZN+DIS+RAD+I(B^2)+MEDV+ INDUS_LSTAT+TAX_RAD + I(LSTAT),
             data=data_csv)
summary(result4)
vif(result4)
```