数理統計 補助資料
～線形回帰分析のためのデータ作成～

2024年度2学期:　月曜1限, 水曜3限
担当教員:　石垣　司

重回帰分析を活用してみよう

背景
- あなたはコンサルタントとして,
  スーパーマーケットチェーンの販売促進に
  関する戦略立案を担うことになった

利用できるデータ
- ID-POSデータ: レジ通過時にポイントカードを提示した顧客の購買履歴データ。「誰が, いつ, 何を, 何個, いくらで」購入したのかが記録されているデータ(1年分を利用)
- 顧客の属性データ: 登録顧客の年齢, 家族人数, 世帯内の高齢者の有無, 世帯内の子供の有無, 自宅から店舗までの所要時間

戦略立案のためにまずは現状を知る
- 顧客属性と購買金額の関係を定量的に把握する

重回帰分析～データの整形 #1

ここで用いるデータは実データを元に授業用に作成した人工データである。
しかし、その分析結果は実データの傾向が反映されている

ID-POSデータ

購買日	購買時間	顧客ID	商品カテゴリコード	商品コード	価格	購買個数
2020.05.15	11.15.01	100001	12321	49000000001	198	3
2020.05.15	11.15.01	100001	10089	49011123400	258	1
2020.05.15	11.16.11	123456	10105	49000067592	154	2
…	…	…	…	…	…	…

顧客属性データ

顧客ID	年齢	家族人数	高齢者の有無	子供の有無	家からの距離
00001	61	3	0	0	15分
00002	40	4	0	1	10分
00003	59	2	0	0	25分
…	…	…	…	…	…

重回帰分析用に加工した購買金額データ

顧客ID	購買金額	年齢	家族人数	高齢者の有無	子供の有無	家からの距離
00001	¥267,120	61	3	0	0	15分
00002	¥156,990	40	4	0	1	10分
00003	¥143,428	59	2	0	0	25分
…	…	…	…	…	…	…
01000	¥84,143	71	2	1	0	5分

重回帰分析～データの整形 #2

重回帰分析用に整形した購買金額データ
- このデータがあれば目的の回帰分析が可能

顧客ID	購買金額	年齢	家族人数	高齢者の有無	子供の有無	家からの距離
00001	¥267,120	61	3	0	0	15分
00002	¥156,990	40	4	0	1	10分
00003	¥143,428	59	2	0	0	25分
…	…	…	…	…	…	…
01000	¥84,143	71	2	1	0	5分

～データと数式の対応表～

目的変数ベクトル $\boldsymbol{y}$　　説明変数行列 X の要素

顧客ID	購買金額	年齢	家族人数	高齢者の有無	子供の有無	家からの距離
$i = 1$	y_1	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
$i = 2$	y_2	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}
$i = 3$	y_3	x_{31}	x_{32}	x_{33}	x_{34}	x_{35}
…	…	…	…	…	…	…
$i = 1000$	y_N	x_{N1}	x_{N2}	x_{N3}	x_{N4}	x_{N5}

説明変数作成の注意：$N > P$

データのサンプルサイズ(N)よりも説明変数の種類(P)が大きい場合，回帰係数は推定できない

- **直感的な説明：** $N = 5, P = 10$ **の回帰分析は，5個のデータで10個の未知パラメータを推定したいということ。そもそもの情報が少なすぎて回帰係数に関する情報を得ることができない**

例：

$$
\begin{aligned}
y_1 &= b_0 + b_1 x_{11} + b_2 x_{12} + \cdots, + b_{10} x_{1,10} + e_1 \\
y_2 &= b_0 + b_1 x_{21} + b_2 x_{22} + \cdots, + b_{10} x_{2,10} + e_2 \\
y_3 &= b_0 + b_1 x_{31} + b_2 x_{32} + \cdots, + b_{10} x_{3,10} + e_3 \\
y_4 &= b_0 + b_1 x_{41} + b_2 x_{42} + \cdots, + b_{10} x_{4,10} + e_4 \\
y_5 &= b_0 + b_1 x_{51} + b_2 x_{52} + \cdots, + b_{10} x_{5,10} + e_5
\end{aligned}
\quad \Rightarrow \text{適切な回帰係数は推定不可能}
$$

対策法1：サンプルサイズを増やす

対策法2：説明変数の種類を減らす

説明変数作成の注意：ダミー変数化

カテゴリ変数はダミー変数に変換する

- **カテゴリ変数：性別や職業などの名義尺度や一部の順序尺度**
- **ダミー変数化しないと，変数間で完全な多重共線性が生じる**

顧客ID	男性	女性
00001	1	0
00002	1	0
00003	0	1
…	…	…
01000	0	1

ダミー変数

顧客ID	男性
00001	1
00002	1
00003	0
…	…
01000	0

or

顧客ID	女性
00001	0
00002	0
00003	1
…	…
01000	1

顧客ID	会社員	学生	それ以外
00001	1	0	0
00002	0	0	1
00003	0	1	0
…	…	…	
01000	1	0	0

顧客ID	会社員	学生
00001	1	0
00002	0	0
00003	0	1
…	…	…
01000	1	0

など

説明変数作成の注意：多重共線性

多重共線性

- **ある2つ以上の説明変数の間に強い相関関係があること**
- **多重共線性があると回帰係数の推定結果が不安定になる**

check!

$$
X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1P} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{i1} & \ddots & x_{iP} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{NP} \end{bmatrix}
\begin{matrix} \\ \\ \\ \\ \\ \| \\ \boldsymbol{x}_1 \end{matrix}
\quad
\begin{matrix} \\ \\ \\ \\ \\ \| \\ \boldsymbol{x}_P \end{matrix}
\quad \Rightarrow
$$

例えば，相関係数 $Corr(\boldsymbol{x}_1, \boldsymbol{x}_p) \approx 1$ ならば，回帰係数の推定結果は信頼できない

対策法：多重共線性が生じる説明変数を取り除く

- VIF(**分散拡大係数：** Variance Inflation Factor)**の値などを参考にする**

  取り除くべき変数のVIFの値の目安として，理論的根拠は無いが，5 や 10 以上がよく用いられる。

分散拡大係数 VIF #1

回帰係数 b_p の分散拡大係数 VIF

$$VIF = \frac{1}{1 - R_p^2}$$

- R_p^2**：回帰モデル** $\boldsymbol{x}_p = X_{-p}\hat{\boldsymbol{r}} + \boldsymbol{e}$ **の決定係数**

$$
\boldsymbol{x}_p = \begin{bmatrix} x_{1p} \\ \vdots \\ x_{Np} \end{bmatrix}, X_{-p} = \begin{bmatrix} 1 & \cdots & x_{1,p-1} & x_{1,p+1} & \cdots & x_{1P} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & \cdots & x_{N,p-1} & x_{N,p-1} & \cdots & x_{NP} \end{bmatrix}
$$

$\hat{\boldsymbol{r}}$: **回帰係数ベクトルの最小2乗推定量**

- R_p^2 **が大きい ⇔ 説明変数** x_p **は** p **以外のその他の説明変数でよく説明できてしまう**

  決定係数を用いることで，複数の説明変数ベクトルの線形結合との多重共線性の強さを測ることができる

分散拡大係数 VIF #2

最小2乗推定量 $\hat{b}_p$ の分散と VIF の関係

$$V[\hat{b}_p] = \frac{V[e_p]}{(N-1)S_{x_p}} \frac{1}{1-R_p^2}$$

S_{x_p}: x_pの標本分散

この関係の導出は省略

- VIF の値に比例して $V[\hat{b}_p]$ が拡大

  回帰係数の検定では第2種の過誤(誤った帰無仮説を棄却できない)が増加してしまう。実証分析では致命的な性質

R_p^2 の値と VIF の値の関係

- $R_p^2 = 0.8 \Leftrightarrow VIF = 5,\ R_p^2 = 0.9 \Leftrightarrow VIF = 10$

例：スーパーマーケットデータでのVIFの計算

```
> vif(Reg)
     Age   Family      Old    Child     Time
2.325950 1.292024 1.669825 1.441859 1.019022
```


重回帰分析の結果

ソフトウェアを利用して重回帰分析の結果を出力

	Estimate (推定値)	Std.Error (標準誤差)	t value (t値)	Pr(>|t|) (p値)
切片 (b_0)	106146	24196	4.39	0.000***
年齢 (b_1)	841	382	2.21	0.028*
家族人数 (b_2)	23170	2602	8.91	0.000***
高齢者の有無 (b_3)	-1063	8202	-0.13	0.897
子供の有無 (b_4)	7941	7633	1.04	0.299
家からの時間(b_5)	-3208	598	-5.37	0.000**
Adjusted R-squared (自由度調整済み決定係数)		0.106		

データ分析の経験則(※理論的根拠はないが，特に知っておいてほしい事項)

- 卒論研究や実務でのデータ分析では，データの取得，整形に労力の 9 割以上を費やすことが多い
- データを整形するためのプログラミングは必須

プログラミング言語「R」

統計分析に特化した言語

- すべて Free
- 初心者にも扱いやすい
- 様々なパッケージが無料公開
- 回帰分析の結果も簡単に出力
- ダウンロード＆インストール

  「R download」でブラウザで検索

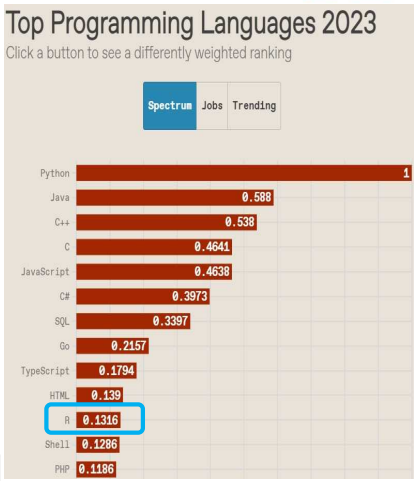
  実行ファイルをクリックするだけで自動的にインストール

Download R-4.4.1 for Windows (82 megabytes, 64 bit)

README on the Windows binary distribution

New features in this version

2024年10月時点の最新バージョン



Top Programming Languages 2023, *IEEE Spectrum*, 29 Aug. 2023

Rによる重回帰分析の結果の出力

重回帰分析の結果の要約

- Reg = lm(Sales~Age+Family+Old+Child+Time, data=Data)
- summary(Reg)

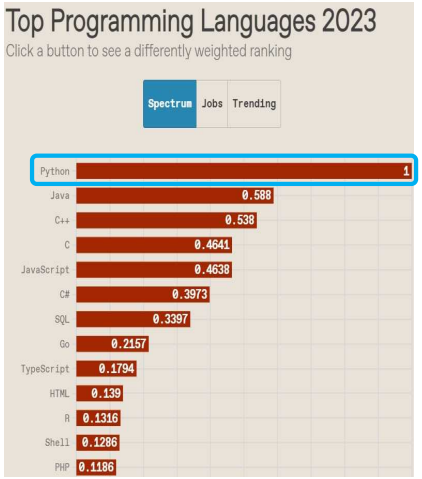
※注意：両側検定によるp値

```
R Console
> Reg = lm(Sales~Age+Family+Old+Child+Time,data=Data)
> summary(Reg)

Call:
lm(formula = Sales ~ Age + Family + Old + Child + Time, data = Data)

Residuals:
    Min      1Q  Median      3Q     Max
-263331  -55158   -3866   58334  207960

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 106146.9    24195.7   4.387 1.27e-05 ***
Age            841.8      381.8   2.205   0.0277 *
Family       23170.6     2601.6   8.906  < 2e-16 ***
Old          -1063.1     8201.6  -0.130   0.8969
Child         7941.5     7633.3   1.040   0.2984
Time         -3208.1      598.0  -5.365 1.01e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 83810 on 994 degrees of freedom
Multiple R-squared:  0.1101,    Adjusted R-squared:  0.1056
F-statistic: 24.59 on 5 and 994 DF,  p-value: < 2.2e-16

>
```

プログラミング言語「Python」

汎用的プログラミング言語

- すべて Free
- 統計分析以外も含む多くの処理を実現可能
- 様々なパッケージが無料公開
- ライブラリ Numpy を利用することで高速な数値計算が可能
- ライブラリ Pandas を利用することで 初心者でも扱いやすいデータ分析環境を利用可能



Top Programming Languages 2023,
IEEE Spectrum, 29 Aug. 2023

本授業で扱う程度のデータ分析はRでもPythonでもどちらでも実行可能

Pythonによる重回帰分析の結果の出力

重回帰分析の結果の要約

```
import statsmodels.api as sm

x = data1.iloc[:,[1,2,3,4,5]]
X1 = sm.add_constant(x)
model1 = sm.OLS(data1["Sales"],X1)
result1 = model1.fit()
result1.summary()
```

OLS Regression Results

Dep. Variable:	Sales	R-squared:	0.110
Model:	OLS	Adj. R-squared:	0.106
Method:	Least Squares	F-statistic:	24.59
Date:	Tue, 03 Oct 2023	Prob (F-statistic):	2.13e-23
Time:	09:17:44	Log-Likelihood:	-12752.
No. Observations:	1000	AIC:	2.552e+04
Df Residuals:	994	BIC:	2.555e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.061e+05	2.42e+04	4.387	0.000	5.87e+04	1.54e+05
Age	841.7752	381.770	2.205	0.028	92.608	1590.942
Family	2.317e+04	2601.630	8.906	0.000	1.81e+04	2.83e+04
Old	-1063.1293	8201.572	-0.130	0.897	-1.72e+04	1.5e+04
Child	7941.4937	7633.264	1.040	0.298	-7037.669	2.29e+04
Time	-3208.0992	598.008	-5.365	0.000	-4381.603	-2034.596

Omnibus:	8.079	Durbin-Watson:	1.983
Prob(Omnibus):	0.018	Jarque-Bera (JB):	5.975
Skew:	0.064	Prob(JB):	0.0504
Kurtosis:	2.644	Cond. No.	498.

当然ではあるのだが、RとPythonの両方の結果がまったく同じであることを確認してほしい

演習問題

完全な多重共線性がある場合, 最小2乗法では回帰係数を推定できない理由を数学的な観点から考えてみよう

1. $N = 3, P = 2$ **のとき, 完全な多重共線性がある説明変数の行列** $X = \begin{bmatrix} 1 & a & a \\ 1 & b & b \\ 1 & c & c \end{bmatrix}, (a \neq b \neq c)$ **を考える。このとき,** $\mathrm{rank}(X)$ **と** $\mathrm{rank}(X^T X)$ **の値を答えなさい**
2. **正方行列** A **に逆行列** A^{-1} **が存在するための条件から,** $X^T X$ **に逆行列が存在するかどうかを答えなさい**
3. **上の** X **を用いて, 最小2乗推定量** $\hat{b}$ **が推定できるかどうかを答えなさい**