経済と社会 補助資料
マーケティングと実証分析1

2024年度2学期:　火曜2限
担当教員:　石垣　司

実証分析とは？

仮説や理論をデータを用いて統計的に検証すること

- 例: 広告費を上げると商品の売り上げは増えるか？
- 例: リフレッシュ休暇をとると生産性は上がるか？

この授業の焦点: 線形回帰モデルの使い方

この授業の目標: 線形回帰モデルを利用した分析結果の数値の意味を正しく解釈できる

	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		

補足事項

線形回帰モデルを利用した実証と予測

- 「実証」のための線形回帰モデル
  - 社会科学での伝統的な実証分析の基礎的ツール
- 「予測」のための線形回帰モデル
  - 予測:過去の構造・パターンを利用し定量的に未知を推計
  - Deep Learning を含む統計・機械学習モデルの基礎
- 実証と予測の両方で線形回帰モデルは中心的基盤

本授業での説明の仕方

- 線形回帰モデル, 統計的検定などの理解には大学初級程度の数学と統計学の知識が必要。しかし, 本授業は全学科目のため, 受講生に対しそれらの知識背景を担保できない。よって, 可能な限りイメージ優先の説明を行う

これ以降のお話の設定

背景: マーケティング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

- 1年分の 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
- 顧客属性データ: 登録顧客の年齢, 家族人数, 世帯内の高齢者の有無, 子供の有無, 自宅から店舗までの所要時間

目標: 戦略立案のためにまずは顧客属性と購買金額の関係を定量的に把握現状を知りたい

データの整形

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
…	…	…	…	…	…

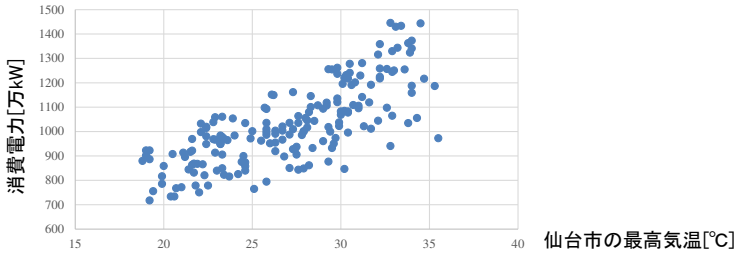
重回帰分析用に加工した購買金額データ

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

回帰分析の前に

基本は「散布図」

- **変数 x（最高気温）と y（消費電力）の相関関係の可視化**
  - **この関係性を利用した予測や実証の手段が回帰分析**



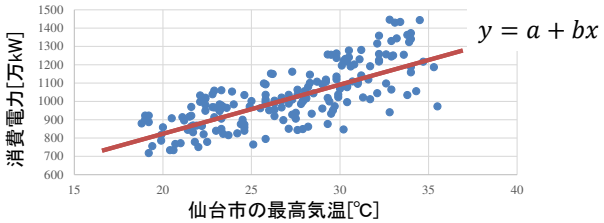
「回帰」とは，目的変数 y の動きを，別の説明変数 x と関数 f で予測したり説明したりすること

東北電力ネットワーク：東北6県・新潟エリアの2020&21年7月1日～9月30日の各日の12時から13時の電力使用量[万kW]
https://setsuden.nw.tohoku-epco.co.jp/download.html
気象庁：2020&21年7月1日～9月30日の各日の仙台市の最高気温
https://www.data.jma.go.jp/obd/stats/etrn/

線形単回帰モデル

関数 f に直線を仮定した説明変数が1つだけの回帰分析のためのモデル

$$y = f(x) = a + bx$$



- **因果関係がある場合は，x が原因，y が結果の表現**
- **データ$\{(x_1, y_1), \cdots, (x_N, y_N)\}$を用いて，散布図の傾向に適合する直線の切片 a と傾き b を推定**
- **切片 a と傾き b が決まれば，目的変数 y を予測できる**

補足：線形単回帰モデルの推定

データから回帰係数 b と切片 a を決定する

- **合理的な基準と手続きに基づいた方法が必要**

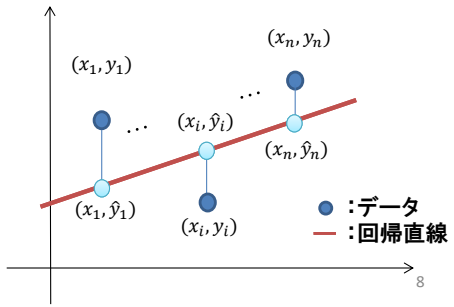
基準：残差平方和の最小化

- **残差** e_i　　$e_i = y_i - \hat{y}_i = y_i - (a + bx_i)$
- **残差平方和**　　$RSS = \sum_{i=n}^{n} e_i^2$

手続き　最小2乗法

最小2乗推定量

$$\hat{b} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2},\ \hat{a} = \bar{y} - \hat{b}\bar{x}$$



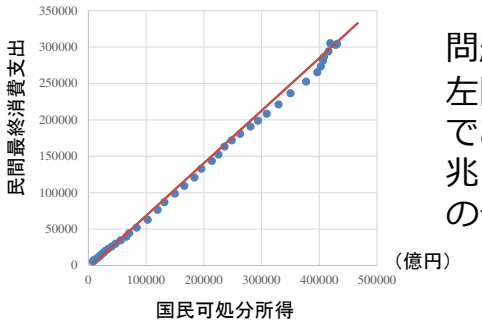
回帰分析と予測

回帰式 $y = \hat{a} + \hat{b}x$ を利用した予測

$\hat{a}$ と $\hat{b}$ はデータ $\{(x_1, y_1), \cdots, (x_N, y_N)\}$ から推定された係数

– $\hat{y}$: 説明変数 x に対する目的変数 y の予測値

$$\hat{y} = \hat{a} + \hat{b}x$$



1955年度～1998年度(1968SNA)
（内閣府　国民経済計算年次推計）

問題
左図の回帰係数は$\hat{a} \cong 0$, $\hat{b} \cong 0.7$ である。国民可処分所得が500兆円のとき，民間最終消費支出の予測値は？

線形重回帰モデル

複数の説明変数による回帰分析

– 目的変数:　変数 y
– 説明変数:　変数 $x_1, x_2, \cdots, x_P$
– データ:　$\{y_i, x_{i1}, x_{i2}, \cdots, x_{iP}\}(i = 1, \cdots, N)$
– 偏回帰係数:　係数 $b_1, b_2, \cdots, b_P$（パラメータ）
– 切片:　係数 b_0（パラメータ）

重回帰モデル

$$y = b_0 + b_1 x_1 +, \cdots, +b_P x_P$$

重回帰モデル(データによる記述)

$$y_i = b_0 + b_1 x_{i1} +, \cdots, +b_P x_{iP} + e_i \ \ (i = 1, \cdots N)$$

線形回帰モデルと実証分析

説明変数が目的変数に影響を与えるか否かの検証

$$y = b_0 + b_1 x_1 +, \cdots, +b_P x_P$$

– 説明変数 x_p が y に影響を与える ⇔ 回帰係数 $b_p \neq 0$
– 説明変数 x_p が y に影響を与えない ⇔ 回帰係数 $b_p = 0$

偏相関係数 b_p の解釈

– p 番目の説明変数 x_p 以外の説明変数の影響を取り除いたときの，目的変数 y と説明変数 x_p の線形関係の傾き

データを用いて偏相関係数 b_p の値を推定

– 推定の方法は非常に重要であるが，本授業では扱わない
– 実際にはソフトウェアを利用して推定値を得る

プログラミング言語「R」

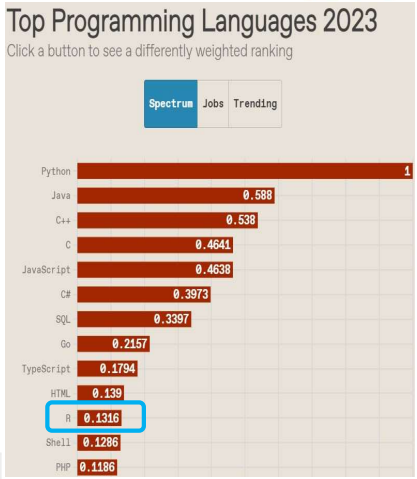
統計分析に特化した言語

– 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

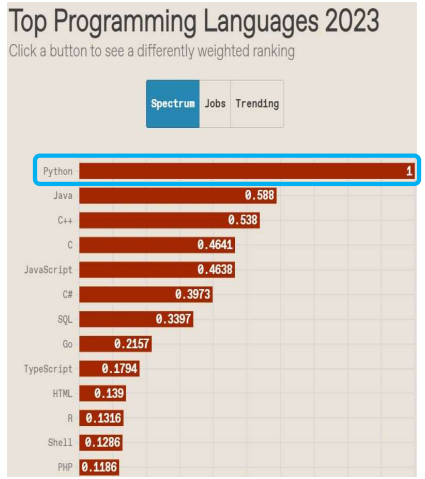
R Console

```
> 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

実証分析と線形回帰モデル

マーケティング・リサーチを含む社会科学分野の実証分析の多くには線形回帰モデルが利用される

説明変数(x)と目的変数(y)の間に線形関係がある場合，統計的な仮説検定により，“関係あり”を主張できる

- 正しい実証のためには，下表の意味を正しく解釈する必要あり

	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		