重回帰分析

M1　魚野 翔太

発表の流れ

- 重回帰分析とは？
- 最小二乗法を用いた回帰方程式の求め方
- 回帰方程式の解釈
- 回帰方程式の精度の指標
- 多重共線性と残差の分析
- SPSSによる重回帰分析

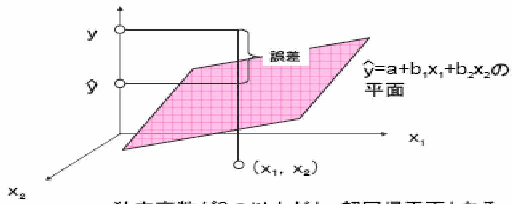
重回帰分析とは？

- 目的変数(従属変数)の変動を2つ以上の説明変数(独立変数,ダミー変数でもよい)の関係式で説明する為に用いる分析法。もしくは、目的変数に強い影響力をもつ説明変数を探すための分析法。

- $y^\wedge = a + b_1x_1 + b_2x_2 + \cdots + b_nx_n$と仮定
- y^ は目的変数の予測値　aは切片　bは偏回帰係数
- 上の線形式をデータにあてはめて最も良く適合する定数項(a, b)を求める
- 以下では説明変数が2つの場合について考える。

3変数の回帰方程式

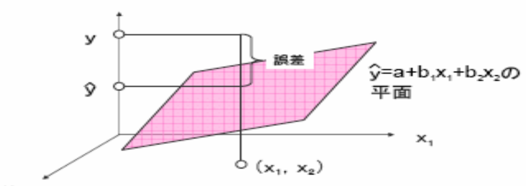
- 3変数の回帰方程式 $y^\wedge = a + b_1x_1 + b_2x_2$は平面を表す(回帰平面)。



独立変数が3つ以上だと，超回帰平面となる

最も適合する回帰式を求めるために

- 最も適合するということは実測値yと予測値y^との誤差(残差 e)が全体として最小になるということ。



昨年度の栗本さんの発表資料から転載

最小二乗法

- 単回帰分析と同様に最小二乗法を用いて $e_i = y_i - y^\wedge = y_i - (a + b_1x_1 + b_2x_2)$の平方和が最小になるような定数項を求める。

- 誤差の平方和$Q = \Sigma[y_i - (a + b_1x_1 + b_2x_2)]^2$

例えば・・・涌井・涌井(2002)より

営業所	投資額(百万) x1	事務員数 x2	生産性(千万) y
1	3	4	28
2	7	5	48
3	4	3	30
4	6	2	30
5	5	1	24
平均	5	3	32

生産性を目的変数、投資額と事務員数を説明変数とする回帰式を求めてみる。

回帰方程式の求め方

- $Q = \Sigma(y_i - y^\wedge)^2$に各データを代入し、展開して整理
- $Q = 5\{a - (32 - 5b_1 - 3b_2)\}^2 + 10\{b_1 + 1/10b_2 - 4\}^2 + 99/100 (b_2 - 140/33)^2 + 3412/11$
- $a-(32 - 5b_1 -3b_2) = 0,\ b_1 + 1/10b_2 - 4 =0, b_2 - 140/33 = 0$ の関係式を満たすa, b, cを計算する。
- $y^\wedge = 46/33 + 118/33x_1 + 140/33x_2$となる。

回帰方程式の解釈

- $a-(32 - 5b_1 -3b_2) = 0$ を変形すると$32 = a + 5b_1 + 3b_2$となる。
- 上記の表を見てみると・・・
- 生産性の平均 = $a + b_1$(投資額の平均) + b_2(事務員数の平均)
- つまり、各変数の平均は回帰方程式を満たす
- $y^- = a + b_1x_1^- + b_2x_2^-$
- 回帰平面は分布の重心(y^-, x_1^-, x_2^-)を通過する。

回帰方程式の解釈

- $y^\wedge = 46/33 + 118/33x_1 + 140/33x_2$
- 投資額x_1を100万増やせば生産性が3576万向上する。
- 説明変数が1単位増加したときに目的変数がどれだけ増加するかを知ることができる。
- 偏回帰係数は他の説明変数が変化しないときにある説明変数の変化によってどれだけ目的変数が影響されるかを示す。

予測値と実測値の関係

- 目的変数 y の平均 = 予測値 $y^\wedge$ の平均
- 残差 e の平均 = 0
- 実測値 y の分散 = 予測値 $y^\wedge$ の分散+残差 e の分散
- が成り立っている。

回帰方程式の予測の精度を考える1

- データがすべて回帰平面上に分布していれば完全な予測が可能
- 一方、回帰平面を中心として球状に分布していれば、そのような場合データを回帰平面で代表させることは無理がある。

回帰方程式の予測の精度を考える1

- $S_y{}^2 = S_{y\wedge}{}^2 + S_e{}^2$
- $S_e{}^2 = 1/(n-1)\Sigma(e_i - e^-)^2$
- $e^- = 0$より
- $=1/(n-1)\Sigma e_i{}^2$
- $=1/(n-1)Q$

- 回帰方程式は予測値の分散が最大になるように定められている。

回帰方程式の予測の精度を考える1

- 実測値の分散$S_y{}^2$は予測値の分散$S_{y\wedge}{}^2$ (回帰方程式では最大)と残差の分散$1/(n-1)Q$(回帰方程式では最小)の和になっている。

- $S_{y\wedge}{}^2$／ $S_y{}^2$は回帰方程式の当てはまりの良さを表現していると考えられる。

- $S_{y\wedge}{}^2$／ $S_y{}^2$を決定係数R^2と呼ぶ($0 \leq R^2 \leq 1$)。

回帰方程式の予測の精度を考える1

- 決定係数が1に近ければより回帰平面に近似する分布を取る。

- 決定係数が0に近ければ球状の分布を取る。

回帰方程式の予測の精度を考える2

- 一般に2つの変数(ここでは予測値と実測値)が密接に関係しているかどうかは相関係数を見れば知ることができる。
- $r_{xy}= S_{xy}/S_xS_y \quad (-1 \leq r_{xy} \leq 1)$
- 2つの説明変数x_1, x_2を適当に合成した変数$w = a + px_1 + qx_2$を作り、実測値yとの相関係数r_{wy}を求める。

- Pとqを適当にあてはめてr_{wy}が最大値を取るような合成変数wを求めると、pとqはそれぞれ回帰式の係数と一致する。

回帰方程式の予測の精度を考える2

- すなわち、回帰方程式は目的変数と最大の相関係数が得られるように説明変数を合成したものと言うことができる。
- 重相関係数$R = r_{wy} = r_{yy\wedge} = S_{yy\wedge}/S_yS_{y\wedge} \quad (-1 \leq |r_{xy} \leq 1)$

- 重相関係数Rが1に近いほど、予測値が実測値に近づいている。
- また、それぞれの定義は異なるが、重相関係数は決定係数の平方根になっている。

決定係数R²の欠点

- 決定係数R^2は回帰方程式のあてはまりのよさを示す。
- しかし、説明変数の数を増やすと決定係数は単純に増加し、見かけ上の精度が上がってしまう。
- この欠点を補うために自由度調整済み決定係数$R^{\wedge 2}$を用いる。
- $R^{\wedge 2} = 1- (n-1)/(n-k-1)\ (1-R^2)$
- nは標本数　kは説明変数の数
- $R^{\wedge 2}$は乱数を説明変数に追加しても値が増加しない。

単位やスケールが異なるとき

- 最初の例
- $y\hat{} = 46/33 + 118/33x_1 + 140/33x_2$
- 投資額x_1を100万円(1単位)増やせば生産性が3576万円向上する
- 1単位を1万にすれば、偏回帰係数は100倍になるので見かけ上大きな影響力があるように見える。
- データを標準化して回帰分析を行えば、(標準)偏回帰係数が目的変数への説明変数の影響度を表現するようになる。

多重共線性の問題

- 独立変数間の相関が高すぎる場合には偏回帰係数の推定量が不安定になる。
- なくてもよい独立変数が説明変数に用いられている。
- 相関の強い独立変数を取り除くか(VIF>10およびCI>15を指標にする)、新しい変数を加えるか、相関する複数の変数を一つの変数に合成するなどの方法をとる必要がある。

残差の分析

- 回帰分析では残差が説明変数と無相関であることを仮定している。
- 残差と説明変数に相関があるなら、その2つをプロットしたグラフには直線的および周期的なパターンが生じる。

- 説明変数の選び方が不適当か線形関係以外の関係がある。
- 変数の変換(対数変換)や他の分析を用いる必要がある。

- 多重共線性と残差の問題がクリアできなければ回帰分析を用いることはできない。

分散分析との違い

- 連続変量を扱うことができる。
- 変数間の交互作用を検討するかどうか
- 変数同士の積を交互作用項として投入することで重回帰分析でも交互作用を検討できる。
- しかし、交互作用の一つのパターンしか検討できない。

SPSSによる重回帰分析

- 例題(Brooks & Meltzoff (2006)を参考に作成したダミーデータ)
- 生後10ヶ月児のどのような能力が生後18ヶ月における言語成績をどれくらい予測できるかということを調べたい。

SPSSによる重回帰分析

- 18ヶ月での言語能力を目的変数、10ヶ月での視線追従、指差し理解、情動理解を説明変数として考える。

- とりあえず強制投入法で分析した結果

変数選択の手順

- 変数増加法　予測の精度を高める変数の順に追加。
- 変数減少法　予測の精度に高めない変数を順に削除。
- ステップワイズ法　いったん追加された変数でも、他の変数を追加したときに予測の精度を高めなかれば削除する(逆も)。
- 全組み合わせ法　説明変数のあらゆる組み合わせの中から最も予測力の高い組み合わせを選択する。

変数間の相関

Correlations

		LANGUAGE	GAZE	POINTING	EMOTION
Pearson Correlation	LANGUAGE	1.000	.574	.175	.293
	GAZE	.574	1.000	.518	.156
	POINTING	.175	.518	1.000	-.004
	EMOTION	.293	.156	-.004	1.000
Sig. (1-tailed)	LANGUAGE	.	.004	.230	.105
	GAZE	.004	.	.010	.256
	POINTING	.230	.010	.	.493
	EMOTION	.105	.256	.493	.
N	LANGUAGE	20	20	20	20
	GAZE	20	20	20	20
	POINTING	20	20	20	20
	EMOTION	20	20	20	20

決定係数と重相関係数

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.623[a]	.388	.273	6.092

a. Predictors: (Constant), EMOTION, POINTING, GAZE

回帰方程式の検定

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	375.913	3	125.304	3.376	.044[a]
	Residual	593.837	16	37.115		
	Total	969.750	19			

a. Predictors: (Constant), EMOTION, POINTING, GAZE

b. Dependent Variable: LANGUAGE

偏相関係数とその検定

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	8.588	6.216		1.382	.186		
	GAZE	2.127	.800	.619	2.659	.017	.707	1.414
	POINTING	-.601	.958	-.144	-.627	.539	.725	1.380
	EMOTION	.718	.728	.196	.986	.339	.966	1.035

a. Dependent Variable: LANGUAGE

多重共線性の診断

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions			
				(Constant)	GAZE	POINTING	EMOTION
1	1	3.809	1.000	.00	.00	.00	.01
	2	.108	5.931	.00	.09	.15	.55
	3	4.944E-02	8.777	.14	.88	.33	.02
	4	3.322E-02	10.707	.85	.02	.52	.42

a. Dependent Variable: LANGUAGE

結果の解釈

- 情動理解と指差し理解の成績は言語能力を予測しない。
- 視線追従の成績は言語能力を予測するのに有効な変数であると言える。
- 視線追従と指差し理解の間には有意な相関が見られる。
- VIFとCIは多重共線性を示していない。

参考文献とHP

- Brooks, R., & Meltzoff, A. N. (2006). The development of gaze following and its relation to language. *Developmental Science*, 8, 6, 535-543.
- 南風原朝和 (2002) 心理統計学の基礎 統合的理解のために　有斐閣
- 森敏昭・吉田寿夫 (1990) 心理学のためのデータ解析テクニカルブック　北大路書房
- 涌井良幸・涌井貞美 (2002) 図解でわかる回帰分析　日本実業出版社
- 栗本達児　単回帰分析と重回帰分析　URL: http://kyoumu.educ.kyoto-u.ac.jp/cogpsy/personal/Kusumi/datasem05/kurimoto.pdf