ロジスティスク回帰分析

2014/4/30

教育学研究科M1

柳岡開地

はじめに

- 統計が苦手な人による統計が苦手な人への説明にしたい (すごーく分かっている人の説明は, 逆に分かりにくい)
- クリティカルな質問には面食らいます
- 自分の研究を材料に, 架空のデータでロジスティク回帰分析を実践してみた (一種の宣伝でもあるのです！)

回帰分析と同じところ

- ロジスティック回帰分析は線形回帰分析 (先週のかわむらくんの発表してくれた) と同様に, 従属変数を $y = ax + b$ と表す (説明変数が多くなれば, $y = ax_1 + bx_2 + c$ となります)

→「説明」や「予測」を目的としている

(ロジスティック回帰分析は, もともと疫学研究において複数個存在する「リスクファクター」を検討するために用いられた)

回帰分析と違うところ

- 何が違う？

線形回帰分析
→ 従属変数が量的変数

ロジスティック回帰分析
→ 従属変数が質的変数
(2値変数で, 賛成・反対, 有・無)
→ 説明変数は連続値でも名義変数でも大丈夫

ロジスティック回帰分析の種類

- 質的変数は何も2値データだけとは限らない

- 今回扱わないが次の3つがある

2値変数の場合 = ロジスティック回帰分析

順序尺度の場合 = 順序ロジット分析

従属変数が3つ以上の場合 = 多項ロジット分析　（中日, 阪神, 巨人)

回帰分析と違うところ

- 別に質的変数でも, 重回帰分析したらよくない？

→ 現にSPSSは結果をはじき出してくれる

う～ん、残念！

→ 重要なのは, 従属変数が正規分布に従うかどうか。質的変数は従わない。

→ 回帰分析はデータの分布に正規分布が仮定されている

ロジスティック回帰分析のいいところ

- 独立変数の尺度, 分布型に対して厳密な仮定をおいていない
- 係数としてオッズ比を求めることができ, 解釈が容易である
- 各対象者につき, 事象の起こる確率を求められる

（欠点も一応)

→モデル構築の判定基準が数種類あり, これを基準にすれば最適な結果が得られると断言できない

対馬 (2010)より引用

ロジスティック回帰分析の理論的なところ

- ロジット変換をなぜするのか

- ロジスティック回帰モデル

なんでロジスティック？

- ロジスティック回帰 (Logistic regression) のロジスティックって何？

- ロジット関数を用いるからしい

→　ロジットとは, 0から1の値をとるpに対して,

$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \log(p) - \log(1-p)$ です

なぜ log が出てくるんだ！高校生以来…

ロジスティック回帰の考え方

- ロジスティック回帰では, 2値データの1が出る正起確率を求めたい。確率なので, 0から1の範囲に限定される

- しかし, y ＝ ax＋b では, yは無限の値をとりえてしまう。そこで, yに細工を加えよう

→ さきほどページの「ロジット変換」を行うことで, 0から1の範囲をとっぱらってしまおうという試み

ロジスティック回帰モデル

- $X = (x_1, x_2, \ldots, x_r)$ という状態のもとで, 現象が発生する条件付き確率を$p(x)$ で表す
- これは次のように表されることが多い

→ $p(X) = \Pr\{\text{発生} \mid x_1, x_2, \ldots, x_r\} = F(x_1, \ldots x_r)$

たとえば, 1個の変数の影響を線形な変数として,

$Z = \beta_0 + \beta_1 x_1$とおく

$$F(Z) = p(X) = \frac{\exp(Z)}{1+\exp(Z)}$$

→ロジスティック関数　丹後 他 (1996)より引用

ロジスティック回帰モデル

$\log\frac{p(X)}{1-p(X)} = \beta_0 + \beta_1 x_1$となる

→右辺の形は重回帰分析によく似ている

さらに, 両辺の指数をとると,

$$\frac{p(X)}{1-p(X)} = \exp(\beta_0 + \beta_1 x_1)$$

この分母に非発生率, 分子に発生率を算出したものをオッズと呼び, 発生確率が非発生確率の何倍かを意味している →普通の比　丹後 他 (1996)より引用

ロジスティック回帰分析の結果を判定する指標

- オッズ比について
- 係数について (最尤法)

オッズ比

- しかし, 実際ロジスティック回帰分析の結果に出てくるのはオッズ比 (比の比)

→オッズ比とはなにか？

A条件とB条件があった場合, オッズ比は…

Aのオッズ÷Bのオッズ= オッズ比 $\exp(b_1)$

→つまり, 要因x_1が1単位増加するに伴って生じるオッズの増大を意味している (割り算することで他の要因が消える)

→オッズではだめなのか？

オッズ比

- 忘れ物があるかどうか

	A君	B君	オッズ
ない	90	99	0.99/0.90 = 1.1
ある	10	1	0.01/0.10 = 0.1

	A君	B君	オッズ
ない	50	55	0.55/0.50 = 1.1
ある	50	45	0.45/0.50 = 0.9

90→99と50→55が同じ比になってしまう…

そこで, オッズ比！

オッズ比

- 忘れ物があるかどうか

	A君	B君	オッズ比
ない	90	99	$\frac{0.99/1-0.99}{0.9/1-0.9} = 11$
ある	10	1	

	A君	B君	オッズ比
ない	50	55	$\frac{0.55/1-0.55}{0.50/1-0.50} = 1.22$
ある	50	45	

90→99の方がオッズ比が高い

ななめ掛けで割り算をしてもよい

オッズ比の信頼区間

- 95%信頼区間

同一の調査, 同一の計算方法を用いた場合, 推定した信頼区間の中に100回中95回入る

ロジスティック回帰分析では, *Wald* 信頼区間

$\exp(\beta \pm 1.96 \times$ 標準誤差$)$　→ βは回帰係数

この95%信頼区間に1を含まなければ, その要因は5%水準で有意, 1を含めば5%水準で有意ではない

最尤法

- 線形回帰分析では, 最小2乗法により係数を求めた。ロジスティック回帰分析では, 別に最尤法という手段で係数の値を求める
- 最尤法とは…の前に尤度って何？
- 尤度とは, 「観測データの下での仮説の尤もらしさ」である。つまり, 観測データが出てしまっている状態で, ある係数の確率分布を当てはめようとする試みなのである

最尤法

- 尤度が最大のときの係数が尤もらしいのは, 観測データの実際の分布に似たような形になるからである。

→ 係数が尤もらしいと観測データの少ないところの確率が低くなり, 観測データの多いところの確率が高くなる

→ よって, 尤度が大きくなる最尤法を用いる

モデルの適合度の評価

- Hosmer-Lemeshow検定
- 正判別率

モデルの適合度

- 適合度を調べるHosmer-Lemeshow検定

標本サンプルを10分の1に分けて, 各グループのモデルの良さを検討する。観測値と予測値の適合を評価するため, X^2検定を行う

- **帰無仮説**：「観測値＝期待値」（ロジスティック回帰モデルはデータに適合する）
- **対立仮説**：「観測値≠期待値」（ロジスティック回帰モデルはデータに適合しない）

→ p <.05 ならばモデルが適合していないことになる

モデルの適合度

- 適合度を調べる判別分割表
- 各対象者のスコアSを算出して, 確率$p(X)$を求める
- $p(X) = 0.5$として判別したとき, 分割表を作成する
- 右下の全体の%が100に近ければ, モデルの適合度は高いと言える。モデルに適合しているかどうかの基準は70%

注意点

- ロジスティック回帰分析にも多重共線性の問題が存在する
- 回帰式に, 相関の高い変数を組み合わせていれたときに, 回帰式が変な値をとる場合が存在する

→ 確認手段として, 相関係数がr >.90となるような相関の高い変数の組み合わせが存在するかがある

→ 線形回帰分析のように, 値を出して調べてくれるところがSPSSにはない

自分の研究とからめてみた！

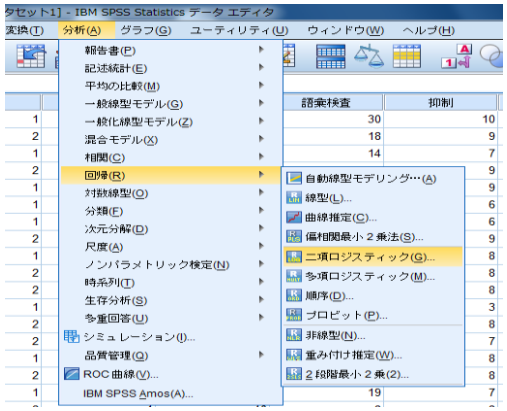
- いつもは当然のようにうまくいくこと（自分の研究は服の着替え) でも,「いつもと異なる」状況に立たされたとき幼児はどうするの？
- 注目したのは「後戻り」をするかしないか。
- では,「いつもと異なる」状況で後戻りができるのは, どうして？そこで, プラニングと実行機能が与える影響について検討してみた

→ 架空のデータだけど, ロジスティック回帰分析だ！

実際にやってみよう！

① SPSSを起動

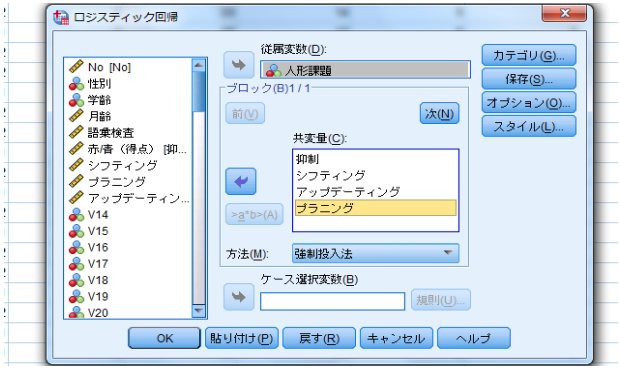
②「分析」→「回帰」→「二項ロジスティック」を選択する



実際にやってみよう！

③「従属変数」のところに人形課題をいれる

④「共変量」のところにプラニング, 抑制, シフティング, アップデーティングをいれる

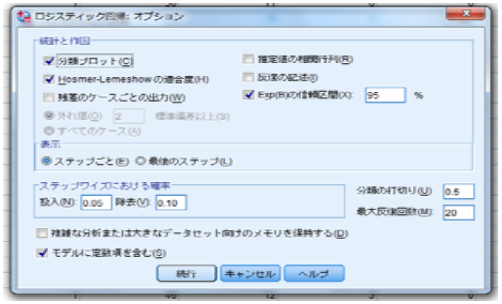


実際にやってみよう！

⑤ 方法は強制投入法で行う

⑥「オプション」をクリック

→「分類プロット」,「Hosmer-Lemeshowの適合度」,「Exp (B) の信頼区間」にチェックし,「続行」



実際にやってみよう！

モデル係数のオムニバス検定

		カイ２乗	df	有意確率
ステップ１	ステップ	57.568	4	.000
	ブロック	57.568	4	.000
	モデル	57.568	4	.000

モデルが $p<.05$ であれば, モデル式が有意であるといえる

実際にやってみよう！

Hosmer と Lemeshow の検定

ステップ	カイ 2 乗	df	有意確率
1	3.551	8	.895

P≧.05 であれば，予測精度が高いことを意味する

実際にやってみよう！

分類テーブル[a]

観測			予測		
			人形課題		
			0	1	正解の割合
ステップ 1	人形課題	0	19	3	86.4
		1	6	58	90.6
	全体のパーセント				89.5

a. 分類値は .500 です

判別の的中率。モデルに適合しているといえそう。

実際にやってみよう！

方程式中の変数

		B	標準誤差	Wald	df	有意確率	Exp(B)	EXP(B) の 95% 信頼区間	
								下限	上限
ステップ 1[a]	抑制	.631	.299	4.456	1	.035	1.880	1.046	3.379
	シフティング	.439	.163	7.271	1	.007	1.552	1.128	2.136
	アップデーティング	-.295	.349	.715	1	.398	.744	.375	1.476
	プラニング	1.764	.867	4.145	1	.042	5.838	1.068	31.917
	定数	-6.593	3.059	4.645	1	.031	.001		

a. ステップ 1: 投入された変数 抑制, シフティング, アップデーティング, プラニング

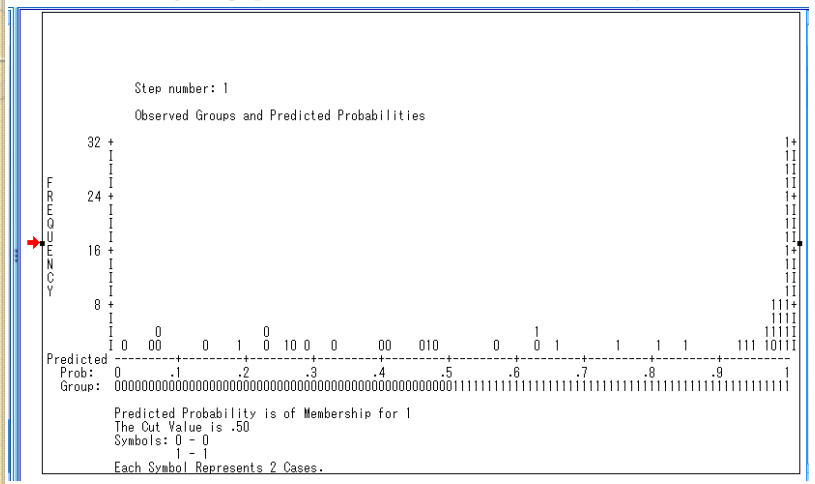
モデル式の係数

定数は無視する

これがオッズ比

オッズ比の信頼区間

実際にやってみよう！



- 分類プロットは、1と0が左右に分かれたグラフの場合うまく予測しているといえる

ちょっと、まった！

- 今までした分析では，参加児の月齢や語彙能力を統制できていない
- プラニングと実行機能が果たす役割を直接検討できたわけではない

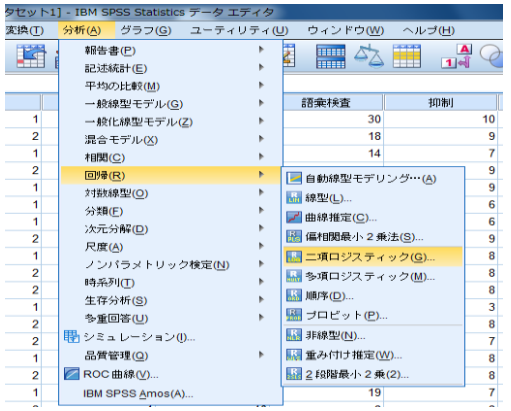
→月齢と語彙能力を統制して，ロジスティック回帰分析を行いたい

→階層的ロジスティック分析

実際にやってみよう！

① SPSSを起動

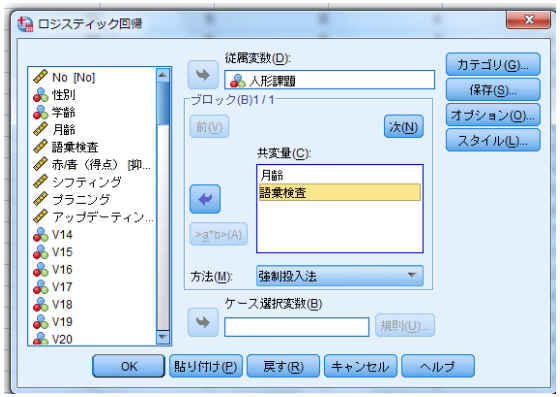
②「分析」→「回帰」→「二項ロジスティック」を選択する



実際にやってみよう！

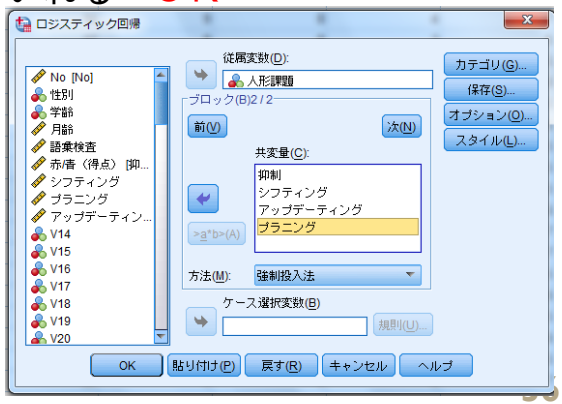
③「従属変数」のところに人形課題をいれる

④「共変量」のところに，まず月齢と語彙検査をいれる



実際にやってみよう！

⑤「次」を押して，ブロック２のところにプラニング，抑制，シフティング，アップデーティングをいれる→ＯＫ



実際にやってみよう！

Hosmer と Lemeshow の検定

ステップ	カイ 2 乗	df	有意確率
1	3.565	8	.894

分類テーブル[a]

観測			予測		
			人形課題		
			0	1	正解の割合
ステップ 1	人形課題	0	19	3	86.4
		1	5	59	92.2
	全体のパーセント				90.7

方程式中の変数

		B	標準誤差	Wald	df	有意確率	Exp(B)	EXP(B) の 95% 信頼区間	
								下限	上限
ステップ 1[a]	月齢	-.103	.101	1.046	1	.306	.902	.741	1.099
	語彙検査	.180	.094	3.644	1	.056	1.198	.995	1.441
	抑制	.610	.334	3.330	1	.068	1.841	.956	3.546
	シフティング	.488	.182	7.156	1	.007	1.628	1.139	2.328
	アップデーティング	-.461	.396	1.356	1	.244	.631	.290	1.370
	プラニング	1.761	.876	4.041	1	.044	5.821	1.045	32.419
	定数	-3.041	4.860	.391	1	.532	.048		

a. ステップ 1: 投入された変数 抑制, シフティング, アップデーティング, プラニング

注意点

- オッズ比の信頼区間
- 今回の架空データをみると, プラニングのオッズ比の信頼区間が異様に大きいことが分かる→なぜか？

可能性①

多重共線性の問題で相関の確認が必要

可能性②

プラニングの1～7点の間に0人のセルがある

→今回は可能性②で, プラニングは独立変数として適切ではないといえる

注意点

- 変数選択法

①強制投入法

・複数の説明変数を同時にモデルに投入する

・独立変数の重要性の順序などの仮説がない場合に使用

②変数増加法

指定した独立変数のうち従属変数に最も強く関連している変数が選ばれ, 以後順番に相関の強い変数が選ばれる

注意点

③変数減少法

指定した独立変数のうち従属変数に対して最も関連が弱く有意でない変数から順番に削除されていく

→②,③には種類がある

変数増加 (減少) 法：尤度比（推奨されている？）

変数増加 (減少) 法：Wald

変数増加 (減少) 法：条件付き

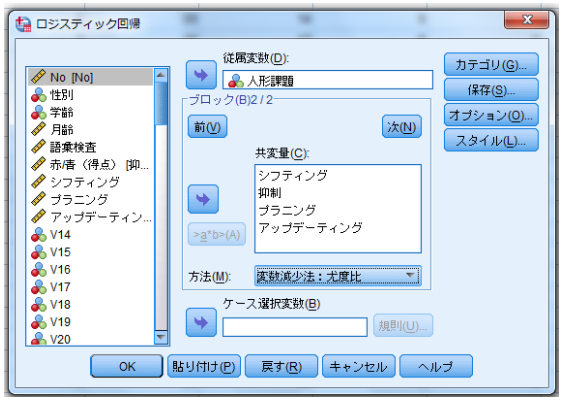
など

変数選択法の選択

- では, 分析の際には強制選択法か変数増加(減少) 法：尤度比の方法で結果が異なるのか
- 実際に比較してみた…

実際にやってみよう！

- 方法のところを変数減少法：尤度比に変えてみる



実際にやってみよう！

Hosmer と Lemeshow の検定

ステップ	カイ 2 乗	df	有意確率
1	3.565	8	.894
2	3.601	8	.891

分類テーブル[a]

観測			予測		
			人形課題		
			0	1	正解の割合
ステップ 1	人形課題	0	19	3	86.4
		1	5	59	92.2
	全体のパーセント				90.7
ステップ 2	人形課題	0	19	3	86.4
		1	5	59	92.2
	全体のパーセント				90.7

a. 分類値は .500 です

モデルは適合しているものの, 数値が少し先ほどとは異なる

実際にやってみよう！

方程式中の変数

		B	標準誤差	Wald	df	有意確率	Exp(B)	EXP(B) の 95% 信頼区間	
								下限	上限
ステップ 1[a]	月齢	-.103	.101	1.046	1	.306	.902	.741	1.099
	語彙検査	.180	.094	3.644	1	.056	1.198	.995	1.441
	シフティング	.488	.182	7.156	1	.007	1.628	1.139	2.328
	抑制	.610	.334	3.330	1	.068	1.841	.956	3.546
	プラニング	1.761	.876	4.041	1	.044	5.821	1.045	32.419
	アップデーティング	-.461	.396	1.356	1	.244	.631	.290	1.370
	定数	-3.041	4.860	.391	1	.532	.048		
ステップ 2[a]	月齢	-.098	.097	1.024	1	.312	.906	.749	1.097
	語彙検査	.164	.091	3.262	1	.071	1.178	.986	1.407
	シフティング	.448	.174	6.654	1	.010	1.566	1.114	2.201
	抑制	.624	.346	3.254	1	.071	1.866	.947	3.677
	プラニング	1.590	.855	3.459	1	.063	4.905	.918	26.208
	定数	-5.701	4.186	1.855	1	.173	.003		

a. ステップ 1: 投入された変数 シフティング, 抑制, プラニング, アップデーティング

ん？結果が違う…

結果の違い

- なんで結果が違うのでしょうか？
- みんなで考えよう！

結果の違い

- ステップ２では, アップデーティングを除いた他の独立変数で, 別の回帰式を作っている
- 結果が違うのは当たり前！
- 今回の架空のデータの結果から言えるのは, 幼児が「いつもと異なる」状況に対応するのにシフティングが重要な役割を果たしている可能性が高いということ

次回の予定

- 近々ではないですが、いつか縦断研究をしたいと思っています
- マルチレベル分析とかほかにも色々とあると思いますが、どれか1つやりたいなと思っています

文献

- 南風原朝和（2002) 心理学統計の基礎有斐閣アルマ
- 平山るみ (2003) ロジスティック回帰分析. http://kyoumu.educ.kyoto-u.ac.jp/cogpsy/personal/Kusumi/datasem03/hirayama.files/frame.htm
- 石村貞夫・謝承泰・久保田基夫 (2001) よく分かる医学・歯学・薬学のための統計解析. 東京図書
- 羅嬉穎 (2008) ロジスティック回帰分析. http://www.educ.kyoto-u.ac.jp/cogpsy/personal/Kusumi/datasem08/logistic.pdf
- 小塩真司 (2007) SPSSとAmosによる心理・調査データ解析—因子分析・共分散構造分析まで—. 東京図書.
- 丹後俊郎・山岡和枝・高木晴良(1996) 統計ライブラリーロジスティック回帰分析—SASを利用した統計解析の実際—. 朝倉出版.
- 対馬栄輝 (2007) SPSSで学ぶ医療系データ解析. 東京図書.
- 対馬栄輝 (2010). 医療系研究論文の読み方・まとめ方—論文のPECOから正しい統計的判断まで. 東京図書