ロジスティック分析の周辺

経済学研究科D1

野口　寛樹

取り上げるもの

- 多項ロジスティック
- 順序ロジスティック
- プロビット
- トービット

はじめに

- 量的変数　＝　重回帰分析
- 二値変数　＝　ロジスティック回帰分析
  プロビット分析
- 順序尺度　＝　順序ロジット回帰分析
- 名義尺度　＝　多項ロジット回帰分析

はじめに2

	被説明変数	当てはめる分布
プロビット	ゼロか1	累積正規分布
ロジット	ゼロか1	ロジスティック曲線
トービット	ゼロか量的データ	累積正規分布

確認

- どこまでが量的変数か
  社会調査では全てがカテゴリー？？
- サンプルサイズ
  - 重回帰では小さいサンプルでも安定
  - ロジスティクなどは係数推定に最尤法を使うため、最低２００、＋１説明変数でサンプルを２０は追加したい (Long,1997)

Long ,J.S(1997)Regression models for categorical and limited dependent variables. Sage Publication

ロジスティック回帰分析

$$\mathrm{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i},$$

特徴：質的な被説明変数

（1 or 0、二項ロジスティック）

分布はロジスティック曲線

多項ロジット回帰

内容

- ３つ以上の値をとるカテゴリー変数を目的変数をしたい

考え方

- 目的変数の中から基準となるベースカテゴリーを決め、説明変数の変化に応じて回答者が、ベースカテゴリーから他のカテゴリーに移る確率を検討する

多項ロジット回帰2

注意

- ベースカテゴリーの選び方には注意が要る。表しているのは、ベースカテゴリーからの変化が有意といえるか、という検討を行っている。変数にあわせて何回かするのが良い場合もある

多項ロジット回帰3

・4つの被説明、説明を3つを想定すると

$$p_{ij} = \exp U_{ij} \; / \; \Sigma_k \exp U_{ik}$$

P_{ij}の推定を最尤推定法で行う。被説明で選んだものをz_iとして、

$$L(a_{1j}, a_{2j}, a_{3j}, c_j) = \Sigma \log \; p_{iz_i}$$

が最大になるように、a_{1j}、a_{2j}、a_{3j}、c_jを求める事になる

順序ロジット回帰分析

- 説明変数が順序尺度の時用いる
- 潜在的で連続的な態度（潜在変数）の存在を仮定し、その値を推定するもの
- 考え方は多項とほぼ同じ

（Stataの本によると、潜在変数のcut pointをきめるところに差がある）

順序ロジット回帰分析2

考え方として、被説明が3つとすると

$$\mathrm{Prob}(Y=0|x) = 1/(1+\exp(-u_1+\Sigma xb))$$

$$\mathrm{Prob}(Y=1|x) = 1/(1+\exp(-u_2+\Sigma xb))-1/(1+\exp(-u_1+\Sigma xb))$$

$$\mathrm{Prob}(Y=2|x) = 1-1/(1+\exp(-u_2+\Sigma xb))$$

b と u は推定するパラメータベクトル、x は説明変数ベクトル

で、考える

プロビット

- ロジスティックとほぼ同じ考え方
- 推計は最尤法
- 推計の当てはまりには

  マクファーデンのR2 (0.4か0.5あれば可)

  割合によるR2 (的中率)

  AIC
- ロジット･プロビットどちらを使うかはお任せ

プロビット2

$$P = \Phi\left(\frac{\log D - \mu}{\sigma}\right) = \Phi(\alpha + \beta \ \log D)$$

Φ は標準正規分布関数

P は発生率, D は変数の量である。

トービット

式としては

- L（a.b. **σ**|xi,yi）=

$$\prod_{y>0} \phi(yi-a-bxi) \prod_{y=0} \Phi(-a-bxi)$$

Φ累積正規分布　φ正規分布

トービット2

- 質的選択と通常の回帰モデルの複合

例：服　買わない場合はゼロだが、買う場合支出額として数値で示される

- 服購入部分は正規分布で、服購入していない場合を累積正規分布で考える
- 以上二つをあわして尤度関数を・・・

実際にプロビットをしてみよう

- 石村貞夫(2005)SPSSによる多変量データ解析の手順第三版　p40 参照

回帰→プロビット

応答度数変数と観測全般 (被説明変数)

共変量 (説明変数)

実際にプロビットをしてみよう

Parameter Estimates

Parameter			Estimate	Std. Error	Z	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
PROBIT[a]	散布時間		.057	.130	.438	.661	-.199	.313
	対数濃度		1.547	.351	4.408	.000	.859	2.235
	Intercept[b]	1	-2.581	.306	-8.428	.000	-2.887	-2.275
		2	-4.077	.740	-5.511	.000	-4.816	-3.337
		3	-3.487	.602	-5.794	.000	-4.089	-2.885

a. PROBIT model: PROBIT(p) = Intercept + BX

b. Corresponds to the grouping variable 薬の種類.

- Zは1.96より大きいことで意味がある

probit(死亡率)=1.547対数濃度＋0.057散布時間－それぞれの薬の係数

実際にプロビットをしてみよう

Chi-Square Tests

		Chi-Square	df[a]	Sig.
PROBIT	Pearson Goodness-of-Fit Test	7.546	9	.580
	Parallelism Test	3.583	2	.167

a. Statistics based on individual cases differ from statistics based on aggregated cases.

仮説は
・求めたプロビットモデルはよく当てはまる
・3つのグループのモデル式の係数は等しい
両方0.05↑。当てはまりはよく、係数は同じでよい

実際にプロビットをしてみよう

- 従属変数の予測

Cell Counts and Residuals

	Number	薬の種類	散布時間	Number of Subjects	Observed Responses	Expected Responses	Residual	Probability
PROBIT	1	1	4.000	50	44	44.746	-.746	.895
	2	1	3.000	49	42	38.082	3.918	.777
	3	1	3.000	46	24	25.347	-1.347	.551
	4	1	2.000	48	16	16.662	-.662	.347
	5	1	1.000	50	6	7.124	-1.124	.142
	6	2	5.000	48	48	47.428	.572	.988
	7	2	5.000	50	47	48.721	-1.721	.974
	8	2	5.000	49	47	45.636	1.364	.931
	9	2	2.000	48	34	36.548	-2.548	.761
	10	2	1.000	48	18	15.485	2.515	.323
	11	3	5.000	50	48	48.129	-.129	.963
	12	3	4.000	46	43	42.386	.614	.921
	13	3	3.000	48	38	39.081	-1.081	.814
	14	3	2.000	46	27	26.400	.600	.574

実際にプロビットをしてみよう

- １番の薬剤で、散布時間＝１で死亡率５０％の場合の薬の濃度とは

Probit(0.5)=0なので

０=1.547対数濃度＋0.057×１ー2.581

に入れてみればでる

参考文献

http://aoki2.si.gunma-u.ac.jp/RiskAssessment/logit.htmlの図など

石村貞夫(2005)SPSSによる多変量データ解析の手順　第三版　東京図書株式会社

山澤成康 (2004) 実践計量経済学入門　日本評論社

石黒格　編 (2008) Stataによる社会調査データの分析　北大路書房