Stataによる多重代入

株式会社　ライトストーン

2020年10月

第5回 多重代入の利用例

- 第4回までは欠損値のある説明変数は連続変数と仮定していた
- 今回は多重代入を行う変数(説明変数)が、二値変数や多項分布にしたがう場合のStataのコマンドを紹介する
- COX比例ハザードモデルにおける多重代入の実行方法を示す

  - キーワード：ロジスティックモデル、多項ロジスティックモデル、負の二項分布、Nelson-Aalen推定量、COX比例ハザードモデル

1.代入コマンド

- 解析モデルの説明変数の種類によって次のような代入コマンドが用意されている
- mi impute intreg
- mi impute logit
- mi impute mlogit
- mi impute nbreg
- mi impute ologit
- mi impute poisson
- mi impute regress
- mi impute truncreg

この3つのコマンドの利用方法を紹介する

mi impute logit

- ロジスティックモデルの説明変数(二値変数)に欠損値がある場合
  - 二値変数hsgradに多重代入を実行する

```
. webuse mheart2,clear
. mi set mlong
. mi misstable summarize
```

Variable	Obs=.	Obs>.	Obs<.	Obs<. Unique values	Min	Max
hsgrad	18		136	2	0	1

高校卒業資格の有無(0,1)に関する変数hsgradに18個の欠損値が存在する

mi impute logit

. mi register imputed hsgrad
. mi impute logit hsgrad attack smokes age bmi female, add(10)
. mi estimate: logit attack smokes age bmi female hsgrad

hsgradには0または1を代入する

attack	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
smokes	1.287062	.366589	3.51	0.000	.5685582	2.005565
age	.0372878	.0155268	2.40	0.016	.0068558	.0677198
bmi	.1248941	.0466484	2.68	0.007	.0334649	.2163233
female	-.1136558	.4210882	-0.27	0.787	-.9389736	.7116621
hsgrad	.4290148	.4241049	1.01	0.312	-.4029937	1.261023
_cons	-6.301285	1.685845	-3.74	0.000	-9.605518	-2.997052

・ 多重代入の結果、hsgradは有意ではないことが分かった

多項ロジスティック

- 解析モデルで利用する説明が、3種類以上の質的情報を持つ場合に多項ロジスティックモデルを利用する
- 最初に多項ロジスティックモデルの仕組みを解説する

多項ロジスティック

- 例えば、日曜日の午後の過ごし方を次の3つの選択肢から選んでもらう
- 1:テレビ番組を見る、2:読書、3:散歩
- 多項ロジスティック回帰モデルは次のようなモデルを利用する

$$\Pr(y = 1) = \frac{e^{X\beta(1)}}{e^{X\beta(1)} + e^{X\beta(2)} + e^{X\beta(3)}}$$

$$\Pr(y = 2) = \frac{e^{X\beta(2)}}{e^{X\beta(1)} + e^{X\beta(2)} + e^{X\beta(3)}}$$

$$\Pr(y = 3) = \frac{e^{X\beta(3)}}{e^{X\beta(1)} + e^{X\beta(2)} + e^{X\beta(3)}}$$

識別

- ここで、

$$\Pr(y = 1) = 0.2, \Pr(y = 2) = 0.5, \Pr(y = 3) = 0.3$$

このような確率を算出するパラメータが

$$\beta(1) = -0.5, \beta(2) = 1, \beta(3) = 3$$

だけでなく、

$$\beta(1) = 1, \beta(2) = -2, \beta(3) = 0.7$$

と複数存在すると、どちらの推定値のセットが適切なのか不明になる。この状態を識別不能と呼ぶ

識別

- そこで任意のパラメータをゼロとして識別可能な状態を実現する
- 例えば、β(1)=0として、β(2)とβ(3)はβ(1)との違いの大きさとして考える(敢えてパラメータを減らす)
- β(1)=0とした時の選択確率は

$$\Pr(y = 1) = \frac{1}{1 + e^{X\beta(2)} + e^{X\beta(3)}}$$

$$\Pr(y = 2) = \frac{e^{X\beta(2)}}{1 + e^{X\beta(2)} + e^{X\beta(3)}}$$

$$\Pr(y = 3) = \frac{e^{X\beta(3)}}{1 + e^{X\beta(2)} + e^{X\beta(3)}}$$

相対確率

- 選択肢1に対する選択肢2の確率の比(相対リスク比)は

$$\frac{\Pr(y=2)}{\Pr(y=1)} = e^{X\beta(2)}$$

相対リスク比には説明変数Xが含まれている。例えば、Xとして年齢を考える
年齢が1単位増えると、相対リスク比は次のようになる

$$\frac{e^{(X+1)\beta(2)}}{e^{X\beta(2)}} = e^{\beta(2)}$$

相対リスク比の限界効果は一定
*話を簡単にするため、定数項は無視した

健康保険の選択

- 分析の目的:人種によって健康保険の選択肢に違いはあるのか?
- 保険の種類:変数insureには3つの値が入っている
- 1.保障型、2.先払い型、3.保険加入なし
- 人種の変数はnowhite(0:White, 2:Black)

健康保険の選択

```
. use https://www.stata-press.com/data/r16/sysdsn1,clear
. tabulate insure nonwhite, chi2 col
```

insure	nonwhite 0	nonwhite 1	Total
Indemnity	251 50.71	43 35.54	294 47.73
Prepaid	208 42.02	69 57.02	277 44.97
Uninsure	36 7.27	9 7.44	45 7.31
Total	495 100.00	121 100.00	616 100.00

```
Pearson chi2(2) =   9.5599   Pr = 0.008
```

カイ二乗検定から保険選択と人種は独立であるとは言えない。つまり、両者間には何らかの関係あることが示唆されている

多項ロジスティックモデル

. mlogit insure nonwhite

ゼロ制約

insure	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Indemnity	(base outcome)					
Prepaid						
nonwhite	.6608212	.2157321	3.06	0.002	.2379942	1.083648
_cons	-.1879149	.0937644	-2.00	0.045	-.3716896	-.0041401
Uninsure						
nonwhite	.3779586	.407589	0.93	0.354	-.4209011	1.176818
_cons	-1.941934	.1782185	-10.90	0.000	-2.291236	-1.592632

多項ロジスティックモデル

- デフォルトでは選択肢の1番にゼロ制約がかかる
- 2番の選択確率は

$$\Pr(y = 2) = \frac{e^{X\beta(2)}}{1+e^{X\beta(2)}+e^{X\beta(3)}}$$

$$X\beta(2) = \beta_0(2) + \beta_1(2)NonWhite$$

- White(nonwhite=0)が2番を選択する確率は、

$$e^{X\beta(1)} = e^{0+0\times 0} = 1$$

$$e^{X\beta(2)} = e^{\beta_0(2)+\beta_1(2)\times 0} = e^{\beta_0(2)} = e^{-0.187}$$

$$e^{X\beta(3)} = e^{\beta_0(3)+\beta_1(3)\times 0} = e^{\beta_0(3)} = e^{-1.941}$$

多項ロジスティックモデル

- よって、

$$\Pr(insure = \text{Prepaid}) = \frac{e^{-0.187}}{1+e^{-0.187}+e^{-1.941}} = 0.42 \qquad \text{(White)}$$

- 同じ要領でNonwhiteについて計算すると、

$$e^{X\beta(1)} = e^{0+0\times1} = 1$$
$$e^{X\beta(2)} = e^{\beta_0(2)+\beta_1(2)\times1} = e^{-0.187+0.66}$$
$$e^{X\beta(3)} = e^{\beta_0(3)+\beta_1(3)\times1} = e^{-1.941+0.377}$$

$$\Pr(insure = \text{Prepaid}) = \frac{e^{-0.187+0.66}}{1+e^{-0.187+0.66}+e^{-1.941+0.377}} = 0.57 \qquad \text{(Nonwhite)}$$

*この結果はスライド12のPrepaidの選択比率に一致している

mlogitによる多重代入例

この例では、解析モデルの説明変数marstatus(婚姻の状態)は3つのカテゴリを示す数値変数である

```
. webuse mheart3,clear
. tabulate marstatus, missing
```

7個の欠損値が存在する

Marital status: single, married, divorced	Freq.	Percent	Cum.
Single	53	34.42	34.42
Married	48	31.17	65.58
Divorced	46	29.87	95.45
.	7	4.55	100.00
Total	154	100.00	

mi impute mlogit

```
. mi set mlong
. mi register imputed marstatus
. mi impute mlogit marstatus attack smokes age bmi female
          hsgrad, add(10)
. mi estimate: logit attack smokes age bmi female hsgrad i.marstatus
```

attack	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
smokes	1.347604	.3734834	3.61	0.000	.6155896	2.079619
age	.0402731	.0159665	2.52	0.012	.0089792	.0715669
bmi	.1246352	.0467078	2.67	0.008	.0330895	.216181
female	.0260006	.4319501	0.06	0.952	-.8206092	.8726104
hsgrad	.0793306	.4143375	0.19	0.848	-.7327561	.8914174
marstatus						
Married	.7621232	.4498884	1.69	0.090	-.1197431	1.64399
Divorced	-.0838154	.4583194	-0.18	0.855	-.9822656	.8146348
_cons	-6.495345	1.708725	-3.80	0.000	-9.844385	-3.146305

負の二項分布

- 回数をモデル化する場合にポアソン回帰モデルを利用することが多い
- しかし、回数がゼロという情報が多い場合には負の二項分布を使う
- 負の二項分布のコマンドnbregを実行すると、ポワソン回帰モデルを用いるべきか否かの仮説検定の結果をStataは報告する
- 今、心臓発作の有無に関する解析モデル(ロジスティック回帰モデル)を推定したいが、説明変数npreg(妊娠の回数)に欠損値が存在し、ゼロとなる標本が一定数存在する

負の二項分布

```
. use https://www.stata-press.com/data/r16/mheartpois,clear
. nbreg npreg attack smokes age bmi hsgrad if female==1
```

npreg	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
attack	.0551929	.4214484	0.13	0.896	-.7708309	.8812166
smokes	.0521987	.4182004	0.12	0.901	-.7674591	.8718565
age	-.0105877	.0174661	-0.61	0.544	-.0448206	.0236452
bmi	.0194787	.0489883	0.40	0.691	-.0765367	.115494
hsgrad	.5338139	.4972872	1.07	0.283	-.4408511	1.508479
_cons	-.0736959	1.551417	-0.05	0.962	-3.114417	2.967025
/lnalpha	-.7956602	.7987311			-2.361144	.769824
alpha	.4512832	.3604539			.0943122	2.159386

LR test of alpha=0: chibar2(01) = 3.00 Prob >= chibar2 = 0.042

*過分散のパラメータαが0に等しい場合、ポワソン分布を利用する

負の二項分布

```
. mi set mlong
. mi register imputed npreg
. mi impute nbreg npreg attack smokes age bmi hsgrad,
          add(20) conditional(if female==1)
```

conditionalオプションを利用して、多重代入は女性のデータ(npreg)に対してだけ実行する

	Observations per *m*			
Variable	Complete	Incomplete	Imputed	Total
npreg	144	10	10	154

推定

. mi estimate: logit attack smokes age bmi female hsgrad npreg

attack	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
smokes	1.244587	.3608022	3.45	0.001	.5374276	1.951746
age	.036776	.0154568	2.38	0.017	.0064812	.0670708
bmi	.1213376	.0460029	2.64	0.008	.0311735	.2115017
female	-.1583678	.5439879	-0.29	0.771	-1.224629	.9078929
hsgrad	.1079131	.4050452	0.27	0.790	-.6859609	.901787
npreg	.0374809	.2700956	0.14	0.890	-.4920491	.567011
_cons	-5.929425	1.642042	-3.61	0.000	-9.147768	-2.711082

npregは有意ではない

COX比例ハザードモデル

- COX比例ハザードモデルにおける欠損値に対して多重代入を実行する
- 代入モデルの説明変数には、解析モデルのアウトカムを含めるのが一般的
- 次に示すCOX比例ハザードモデルではアウトカムはどうすれば、よいだろうか?

$$h(t) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \cdots)$$

- Stata Pressから出版されているAn Introduction to Survival Analysis Using Stata, Revised Third Editionの例題を紹介する

Nelson–Aalen推定量

- White and Royston (2009)は累積ハザードのNelson-Aalen 推定量をアウトカムとして利用することを提案している

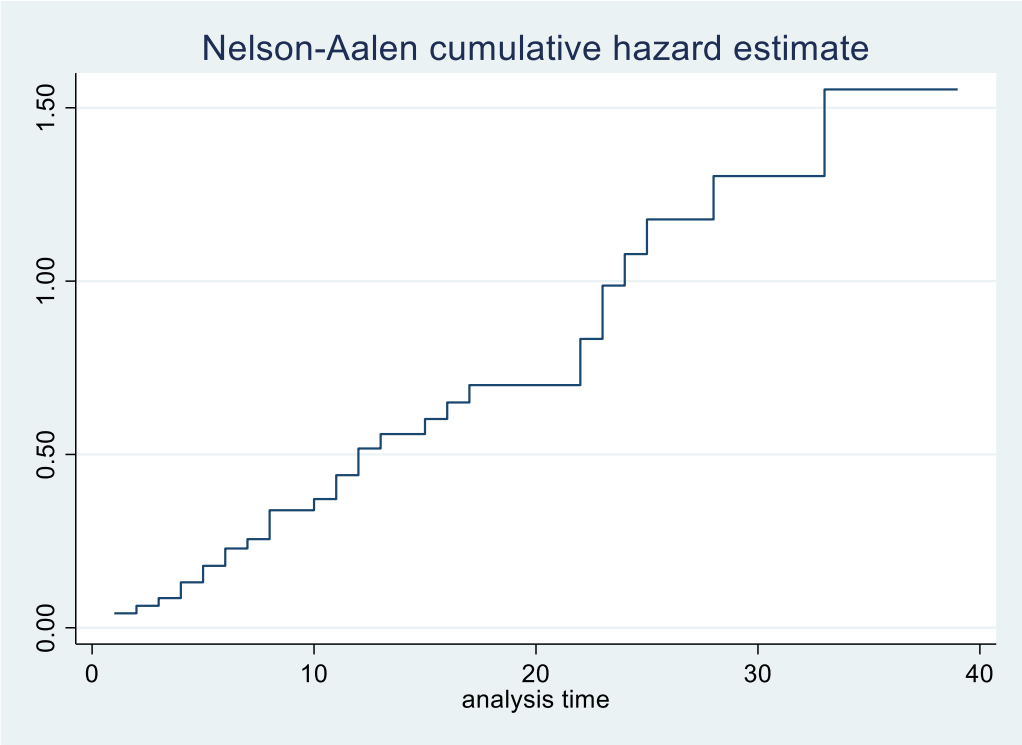
$$\hat{H}(t) = \sum_{j|t_j \le t} \frac{d_j}{n_j}$$

時点tjが到達する直前までのアットリスクのサンプル数をnj, そして時点tjにおけるイベント発生数をdjとする

Nelson–Aalen推定量

```
. use http://www.stata-press.com/data/cgm3r/cancer,clear
. stset
. su,sep(4)
. sts graph,cumhaz
. sts gen cumhaz=na
```

代入モデルで利用する累積ハザードをグラフで確認し、最後にcumhazという名前で取り出す



多重代入

- 欠損値のある変数ageに対して線形回帰モデルを利用して多重代入を実行する

. mi set mlong
. mi register imputed age
. mi impute reg age drug died cumhaz,add(30) rseed(10)

```
Univariate imputation                          Imputations =       30
Linear regression                                    added =       30
Imputed: m=1 through m=30                          updated =        0
```

Variable	Observations per *m*			
	Complete	Incomplete	Imputed	Total
age	40	8	8	48

```
(complete + incomplete = total; imputed is the minimum across m
 of the number of filled-in observations.)
```

推定

. mi est:stcox drug age
. mi est,hr

第3回説明した項目を復習する
Average RVI:欠損値の存在による大きくなる標準誤差の割合。パラメータが2つあるのでその平均を見る

```
Imputations        =          30
Number of obs      =          48
Average RVI        =      0.2004
Largest FMI        =      0.2735
DF:      min       =      398.33
         avg       =   2,832.57
         max       =   5,266.80
F(   2, 1831.9)    =       12.33
Prob > F           =      0.0000
```

Largest FMI:欠損した情報の割合

変数は2つだけなので
398.33はageの自由度、
5,266.8はdrugの自由度と
なる

推定

_t	Haz. Ratio	Std. Err.	t	P>|t|	[95% Conf.	Interval]
drug	.0933812	.045516	-4.86	0.000	.0359145	.2427998
age	1.157173	.0540179	3.13	0.002	1.055704	1.268394

- stcoxコマンドでCOX比例ハザードモデルを推定する
- MIの場合は係数推定値を出力するので、もう一度、hrオプションだけを利用して推定する
- MIは1)欠損値が存在し、情報が完全でないこと、2)代入回数Mが有限であることから、検定統計量はzでなく、t統計量を利用する

まとめ

- 欠損値のある変数が、連続変数でない場合のStataコマンド(代入モデル用)を紹介した
- COX比例ハザードモデルの説明変数おける欠損値への多重代入の実行方法(White and Royston 2009)を解説した

参考文献

- ALLISON, P.D. (2002). Missing Data. Sage Publications
- Cleves, M.A., W.Gould, and Y.V. Marchenko. (2016). An Introduction to Survival Analysis Using Stata, Revised Third Edition. Stata Press
- STATA MULTIPLE IMPUTATION REFERENCE MANUAL RELEASE 16

研究者向けの統計解析ソフトウェア Stata® 評価版のお申込み

機能制限なし！30日間無料で使える評価版でまずはお試しください！ https://www.lightstone.co.jp/stata/evaluate.html

※学生の方への評価版の提供はございません。大学の講義で使用するなど単位取得に関わる形でのご利用となる場合は学生版のご購入をご検討ください。

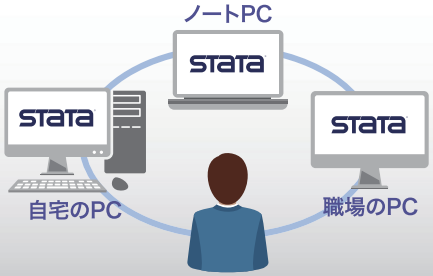
ライセンスはサブスクリプションがおすすめ！

サブスクリプションライセンスを選ぶメリット

- 常にStataの最新バージョンが利用できます
- 技術サポートの対象となります
- 毎年の経費としてStataを導入できます
- 利用者の増減に柔軟に対応可能
- 初期導入費を抑えたい場合にもおすすめ

シングルライセンス（サブスクリプション）

職場のPC・ノートPC・自宅のPC等、個人所有のPCにインストール可能です。
同時に使用できるのは1台のみとなります。



マルチユーザライセンス（サブスクリプション） 2ユーザ以上

ボリュームライセンス

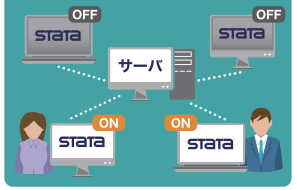
シングルライセンスを複数人でまとめてご購入いただく際のボリュームディスカウント商品です。

同時起動ライセンス

ご利用環境により下記3通りの運用方法からご選択いただけます。

PC毎にスタンドアロンインストールをして運用
例:2ライセンスの場合

クライアント＆サーバ環境によるネットワーク運用
例:2ライセンスの場合



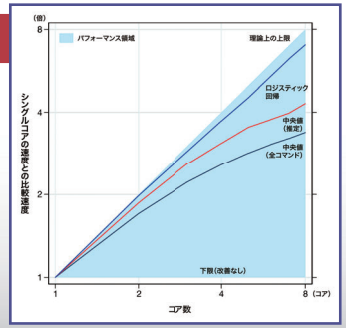
1台の仮想マシンにユーザがアクセスする運用
例:2ライセンスの場合

Stata／MPの演算能力

データ処理時間の削減ならMP

Stata MPはPCの持つマルチコアの特性を活かして、処理を分散・並列する機能を備えます。およそ85%以上のコマンドで処理速度が向上し、コア数に応じた計算時間の短縮が期待できます。
コマンドごとの処理速度の向上の度合いについては以下のページや資料をご覧ください。

コア数	全てのコマンド	推定コマンド	ロジスティック回帰
2	1.7倍	1.8倍	1.9倍
4	2.6倍	3.1倍	3.8倍
8	3.3倍	4.2倍	6.8倍



弊社Webページ内の『Stata／MP』
https://www.lightstone.co.jp/stata/statamp.html

株式会社 ライトストーン
〒101-0031
東京都千代田区東神田 2-5-12 龍角散ビル 7F
TEL：03-3864-5211
FAX：03-3865-0050
e-mail： sales@lightstone.co.jp
WEB： https://www.lightstone.co.jp/