Stataによる多重代入

株式会社　ライトストーン

2020年7月

はじめに

- 第１回では多重代入のイメージを掴むために、大切なキーワードの解説と、Stataによる操作を体験した
- 欠損値に対する対応方法は多重代入以外にも存在する
- 今回は広い視野から、欠損値に対する対応方法について見ていく

第2回 欠損値への対応方法

1.リストワイズ削除

2.最尤法

3.EMアルゴリズム

4.多重代入

- キーワード：周辺分布、条件付き期待値、不確実性、分割表、モノトーンパターン、尤度

1.リストワイズ削除

- Excelのような画面で考えた場合、欠損値のある行をまるごと削除する
- MCARの時には適切である
- 標本はオリジナルデータからのランダムサンプルとなる
- この時、推定値にバイアスは生じない
- データが少なくなる分、標準誤差はやや大きくなるが、その他の統計量は問題ない
- MARの場合、リストワイズ削除にはバイアスが生じる

*MCARとMARに関する解説は第1回にあり

ペアワイズ削除

- 2つの変数X,Yが関心のある変数とする
- XとYのどちらか一方、または、両方が欠損しているデータを削除する
- MCARの時、ペアワイズ削除は一致推定量を提供する
- MARの場合、ペアワイズ削除にもバイアスが生じる

2.最尤法

- 回帰モデルの推定に利用する手法として最小二乗法(OLS)と最尤法(ML)がある
- どちらの方法も被説明変数や説明変数に欠損値が存在する場合、当該行は標本から除外する(ペアワイズ削除)
- ただし、MLには欠損値に対応した手法がある
- 最尤法とは次に示す尤度関数を最大化するパラメータを求めるための手法

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

最尤法

- 最尤法の特徴…大標本であるほど、次に示す3つの特徴を活かすことができる
  - 一致性(不偏性)
  - 漸近的効率性(標準誤差)
  - 漸近的正規性(有意性検定)

最尤法と欠損値

- [仮定]データの欠損メカニズムがignorable(モデルパラメータと、欠損データのメカニズムを支配するパラメータが異なる)であり、MARとする

$$\Pr(Y \text{ is missing}|Y, X) = \Pr(Y \text{ is missing}|X)$$

- 2つの変数XとYについてn個のデータを集めた
- 最初のm個は両方観測できたが、Xのn-m個は観測できなかった
- 2つの変数の同時分布をf(x,y|θ)とし、離散型の変数であるとする

周辺分布

- yの周辺分布

$$h(y|\theta) = \sum_x f(x,y|\theta)$$

- 周辺分布を利用した尤度は、欠損値がモノトーンパターンであるとき、

$$L(\theta) = \prod_{i=1}^{m} f(x_i,y_i|\theta) \prod_{i=m+1}^{n} h(y|\theta)$$

- ここでXとYの同時分布について次の関係を利用する

$$g(x|y;\lambda) = \frac{f(x_i,y_i|\theta)}{h(y|\theta)}$$

*モノトーンパターンの詳細は多重代入の項で解説する

最尤法と欠損値

$$L(\theta,\lambda) = \prod_{i=1}^{m} g(x|y;\lambda)h(y|\theta) \prod_{i=m+1}^{n} h(y|\theta)$$

$$= \prod_{i=1}^{m} g(x|y;\lambda) \prod_{i=1}^{n} h(y|\theta)$$

Xはm個のデータで条件付き分布
を最大化するパラメータλを探す

Yはn個のデータで周辺分布を
最大化するパラメータθを探す

- 欠損値がモノトーンパターンの時、個別に尤度を最大化することを考えればよい

分割表の尤度

- 欠損値の情報を活かす周辺分布の利用方法について考える
- 100人を対象にある調査を行った
- XとYは二値の値を取り、100人はいづれかの選択肢を取る
- しかし、20人分の答えが欠損している。内訳はY=1と答えたもののXが不明な人が15人、Y=2でありながら不明な人が5人

	$Y=1$	$Y=2$
$X=1$	10	30
$X=2$	20	20

分割表の尤度

- 欠損値を無視した尤度は(n=80)

	$Y=1$	$Y=2$
$X=1$	10	30
$X=2$	20	20

$$\hat{p}_{ij} = \frac{n_{ij}}{n}$$

$$L = (p_{11})^{10}(p_{12})^{30}(p_{21})^{20}(p_{22})^{20}$$

- しかし、欠損値の存在を考慮すると、

$$L = (p_{11})^{10}(p_{12})^{30}(p_{21})^{20}(p_{22})^{20}(p_{11}+p_{21})^{15}(p_{12}+p_{22})^{5}$$

周辺分布の利用

- ここでは欠損値を考慮してp_{ij}は次のように周辺分布の情報を利用して計算する

$$\hat{p}_{ij} = \hat{p}(X = i|Y = j)\hat{p}(Y = j)$$

- つまり、

$$\hat{p}_{11} = \left(\frac{10}{30}\right)\left(\frac{30+15}{100}\right) = 0.15$$

$$\hat{p}_{21} = \left(\frac{20}{30}\right)\left(\frac{30+15}{100}\right) = 0.30$$

$$\hat{p}_{12} = \left(\frac{30}{50}\right)\left(\frac{50+5}{100}\right) = 0.33$$

$$\hat{p}_{22} = \left(\frac{20}{50}\right)\left(\frac{50+5}{100}\right) = 0.22$$

周辺分布の利用

- Stataの構造方程式モデリング(SEM)には、モデル推定の選択肢として「欠損値を考慮した最尤法」(MLMV, Maximum Likelihood with Missing Value)という選択肢がある
- これはまさに周辺分布を利用した最尤推定の手法である
- モデルで利用する変数に正規性を仮定している(多変量正規分布)
- 正規分布に従わない二値変数などでも利用可能

3.EMアルゴリズム

- 欠損値の発生メカニズムがモノトーンパターンに従っていない場合はどう対応する?
- EM(Expectation-Maximization)アルゴリズム(Dempster, Laird and Rubin, 1977)を利用する
- 条件付き期待値を欠損値に代入し、最尤法推定の反復計算により目的のパラメータを得る

EMアルゴリズム

- モデルで利用する変数セットをXとする

$$X^T = (Y^T, Z^T)$$

- Yは完全データ、Zには欠損があるとする。密度関数は次のように書くことにする

$$f^C(x|\theta) = f^C(y, z|\theta)$$

- yの確率密度関数は

$$f(y|\theta) = \int f^C(x|\theta)dz = \int f^C(y, z|\theta)dz$$

尤度関数

- 仮に完全なデータが得られていれば、対数尤度関数は

$$l^C(\theta, x) = \log L^C(\theta, x) = \log f^C(x, \theta)$$

- Zの欠損値に条件付き期待値を代入して、この対数尤度関数を利用する

$$E_{\theta^{(k)}}[Z|Y = y]$$

- kは繰り返し回数のk番目を意味する。その時のパラメータがθ^(k)

$$\log f(y|\theta)$$

繰り返し計算

- 対数尤度の条件付き期待値を使って擬似完全データx*を作成し(Eステップ)、次の対数尤度を最大化し(Mステップ)、k+1番目のパラメータを得る

$$l^C(\theta^{(k)}, x^*)$$

- この繰り返し計算を行うことで次のパラメータ系列を得る

$$\{\theta^{(k)}\}$$

- この系列は収束し、観測データに基づく次の尤度を最大化するパラメータと一致する

EMアルゴリズム

- EMアルゴリズムは欠損メカニズムがモノトーンに従わない場合に、最尤法を使って推定量を求める手法
- もちろん、モノトーンに従わない時の統計手法はEMアルゴリズム以外にもある

4.多重代入

- 欠損値の発生メカニズムがMARであるとき、MIを正しく実行すると、パラメータは次の性質を持つ
  - 一致性
  - 漸近的効率性
  - 正規性
- つまり、好ましい検定統計量(t値やz値)を得る
- もちろん、誤った使い方をすると上記の性質は得られないし、不適切な推定値を報告してしまう
- 乱数を利用するので、結果にばらつきが生じる
- しかし、非線形モデルも含めて、様々な計量モデルに対応している

モノトーンパターン

- 欠損パターンに応じて多重代入の方法は異なる
  - 欠損値がモノトーンパターンであるとき、多重代入は計算上、簡単に実行できる(尤度関数の取り扱いが簡単)
  - 今、手元に(X,Y,Z)の3つの変数が4行分存在すると考える

X	Y	Z
1	1	1
0	0	1
0	1	1
0	1	1

行と列を並べ替える

Z	Y	X
1	1	1
1	1	0
1	1	0
1	0	0

モノトーンパターン

  - 1はデータが存在することを示し、0は欠損を示す

モノトーンパターン

- 次のようなデータセットはモノトーンパターンにならない

X	Y	Z
1	1	1
1	1	0
0	1	0
1	0	0

  - モノトーンパターンの場合、複数の変数に欠損が存在しても、一変数づつ独立に代入処理を行うことができる
  - Stataはモノトーン用の代入コマンドを実行すると、モノトーンであることを自動的にチェックする

多重代入

- 二変量正規分布に従う変数XとYの相関係数をMIで推定する
- 数値実験Alison(2002)の流れ
  - 10,000個の乱数を用いて相関係数0.4の変数を作成する
  - 変数Xの半数を削除して欠損値とする
  - 残っている5,000のデータで単純回帰分析を実行し、欠損部分に5,000個の予測値を作成する。そうすると、相関係数が強くなりすぎる
  - 不確実性として単純回帰の残差の1標準偏差を利用した正規分布の乱数を理論値に加えて、再度、予測値を作成する

基礎知識

- 二変量正規分布

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(x-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_x)}{\sigma_x\sigma_y}\right]\right)$$

- ただし、普通のソフトには一変量正規分布の乱数機能しかないので、先にxの乱数を作成し、次に示す条件付き分布を利用する

$$y|x \sim N(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x), (1-\rho^2)\sigma_y^2)$$

- イメージとしてはxに対応してyの平均を動かしながら、乱数を作る
- 標準正規分布を仮定するので、

$$\mu_x = \mu_y = 0, \sigma_x = \sigma_y = 1, \rho = 0.4$$

相関係数と単純なMI

- doファイル(palison_p28_2.do)の内容

```
*相関0.4の変数XとYを作成する
*XYはともに標準正規分布とする

clear
set obs 10000
set seed 1

*最初に変数xの乱数を作成する
gen double x=0
forvalues i = 1(1)10000 {
        replace x = rnormal() in `i'
    }
```

相関係数と単純なMI

```
*条件付き分布を利用してyの乱数を作成する
gen double y=0

forvalues j = 1(1)10000 {
        replace y = sqrt(1-0.4^2)*rnormal()+0.4*x in `j'
    }

corr x y
*rho=0.4087

*5000個の欠損値x2をつくる
gen double x2=x
replace x2=. if _n>5000

reg x2 y
predict double x2_hat
predict u,resid
```

相関係数と単純なMI

```
*5000個を代入したx3をつくる
gen double x3=x2
replace x3=x2_hat if _n>5000

*相関が強すぎる
*rho=0.543
corr x3 y

*残差の標準偏差を調べる
su u
gen double x4=x3
replace x4=x3+r(sd)*rnormal() if _n>5000
*残差の不確実性を利用する
corr x4 y
*rho=0.4143
```

不確実性

- 単純回帰分析でXのフィット値を利用すると、相関係数を過大評価してしまった(r=0.543)
- 不確実性として残差の1標準偏差を利用したところ、ほぼ設定値(r=0.4)に近い値を得た
- 多重代入とは、これと同じ要領でMセットのXYからM個のrを求めてその平均をとり、目的の推定量とする(Rubin, 1987)
- 標準誤差はk番目の相関係数r_kの標準誤差をs_kとすると、

$$S.E.(\bar{r}) = \sqrt{\frac{1}{M}\sum_k s_k^2 + \left(1+\frac{1}{M}\right)\left(\frac{1}{M-1}\right)\sum_k (r_k - \bar{r})^2}$$

多重代入

- このように、多重代入はシミュレーションで作成した推定量の不確実性を利用するというのが基本的な考え方
- 第1回のmi imupte regressの分析例では、パラメータの確率分布という不確実性を利用した
- 欠損値の発生メカニズムがMARであることを仮定していることに注意
- EMアルゴリズムは条件付き期待値を利用した繰り返し計算の収束というアプローチであり、MIとは異なる考え方を利用している

まとめ

- 欠損値の対応方法として一般的に利用される方法を紹介した
- モノトーンパターンの時、最尤法推定では周辺分布を利用すれば欠損が生じているデータ行の情報も利用できる
- モノトーンパターンの有無に関係なく、欠損値の条件付き期待値が解析的に表現できればEMアルゴリズムで欠損値問題に対応できる
- 同じくモノトーンパターンの有無に関係なく、MIは代入モデルによるシミュレーションで複数の推定値を作成し、それらを使って目的のパラメータを求める

参考文献

- ALLISON, P.D. (2002). Missing Data. Sage Publications
- 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。

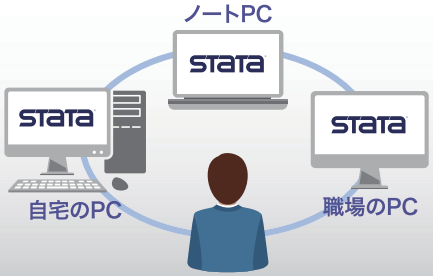
ライセンスはサブスクリプションがおすすめ！

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

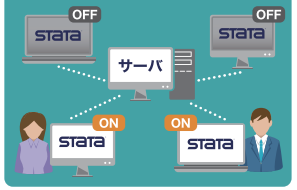
シングルライセンスを複数人でまとめてご購入いただく際のボリュームディスカウント商品です。

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



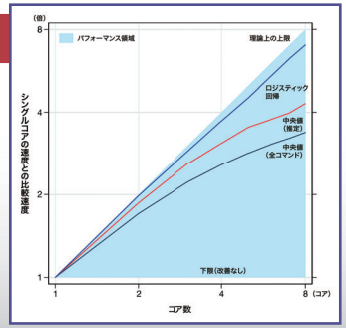
1台の仮想マシンにユーザがアクセスする運用
例:2ライセンスの場合

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/