Stataによる多重代入

株式会社　ライトストーン

2020年6月

2020年8月修正

多重代入

- 欠損値の影響でモデルのパラメータにバイアスが生じていることが心配される場合に用いる統計手法の一つ
- 絶対的な方法は存在しないので、いくつかの方法を試して感度分析まで実行する「時間の余裕」が必要
- 多重代入の手法の中にはベイズ統計の考え方を利用するものがある(ベイズ統計の詳細は「Stataで学ぶベイズ統計」を参照のこと)

構成

- 欠損値に対応する多重代入の基礎知識とStataでの操作方法を5回に分けて学ぶ
  - 第1回：イメージをつかむ
  - 第2回：欠損値への対応方法
  - 第3回：多変量正規分布の利用
  - 第4回：正規性が仮定できない時
  - 第5回：多重代入の利用例

第1回 イメージをつかむ

- Stataの多重代入のPDFマニュアル[MI]にある例題–Intro substantive–を使って多重代入の概略をつかむ
- 統計や数式の説明は最小限に抑え、大きな流れを理解する
  - キーワード: ignorable、MAR、MCAR、MNAR、解析モデル、事後予測分布、代入モデル

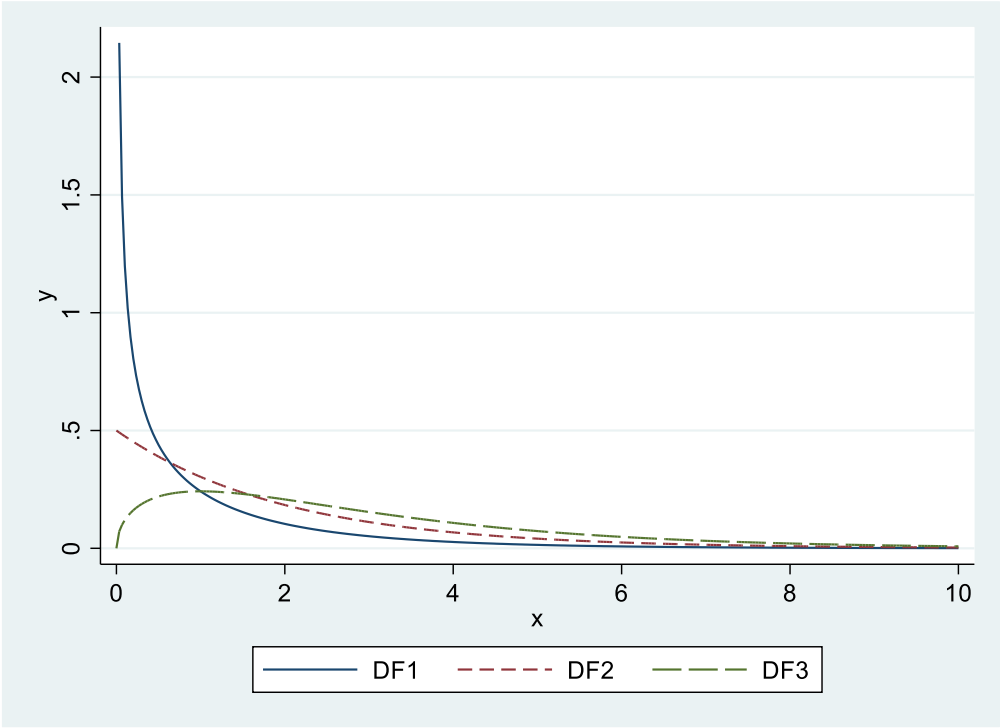
カイ二乗分布

- 後述するStataマニュアルの例題でカイ二乗分布を利用する場面があるので、最初にグラフを使ってイメージを作っておく
  - 正規分布に従う変数Xのn個の標本分散をS2、母分散をσ2とすると、次に示す検定統計量は自由度n−1のカイ二乗分布に従う

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

密度関数

```
twoway function y=chi2den(1,x),range(0 5) || ///
        function y=chi2den(2,x),range(0 5) || ///
        function y=chi2den(3,x),range(0 5)
```



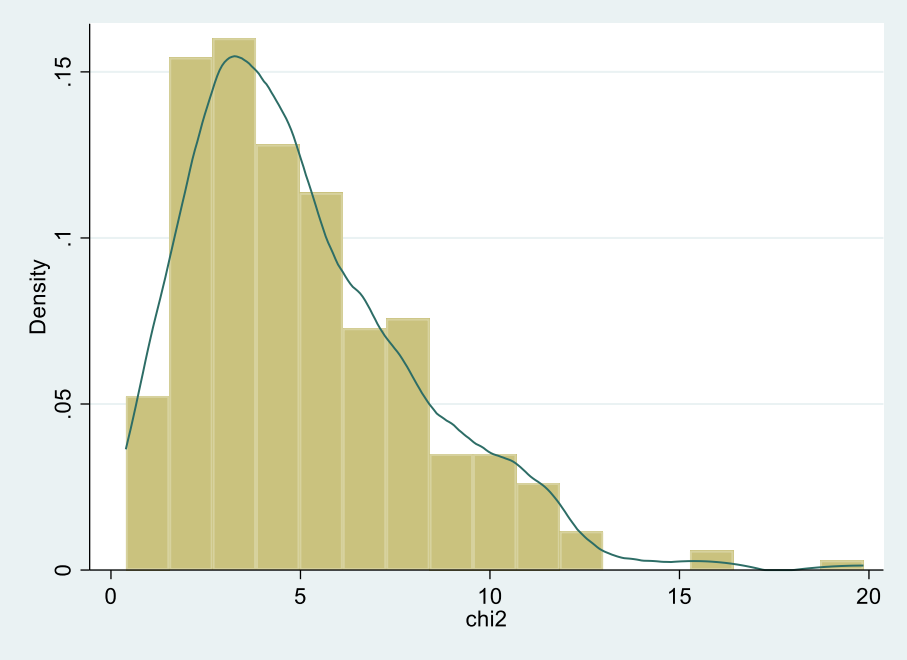
自由度1,2,3の3種類の密度関数プロット

Doファイルを使って実行します

カイ二乗分布による乱数作成

- 自由度5のカイ二乗分布による乱数作成

```
. set obs 300
. set seed 123
. gen chi2= rchi2(5)
. histogram chi2, kdensity
```



作成されるのは、横軸に相当するカイ二乗値

*イメージ作りはここまで

ロジスティック回帰

- 心臓発作の発症に関する仮想データを用いてロジスティック回帰を実行し、リスク要因を考察する

. webuse mheart0,clear
. des

```
  obs:           154                          Fictional heart attack data; bmi missing
 vars:             9                          6 Jun 2020 10:39
-------------------------------------------------------------------------------
                storage   display    value
variable name     type    format     label      variable label
-------------------------------------------------------------------------------
attack          byte    %9.0g                 Outcome (heart attack)
smokes          byte    %9.0g                 Current smoker
age             float   %9.0g                 Age, in years
bmi             float   %9.0g                 Body Mass Index, kg/m^2
female          byte    %9.0g                 Gender
hsgrad          byte    %9.0g                 High school graduate
marstatus       byte    %9.0g      mar        Marital status: single, married, divorced
alcohol         byte    %24.0g     alc        Alcohol consumption: none, <2 drinks/day, >=2
                                                drinks/day
hightar         byte    %9.0g                 Smokes high tar cigarettes
```

解析モデル

. logit attack smokes age bmi hsgrad female

```
Logistic regression                               Number of obs     =        132
                                                  LR chi2(5)        =      24.03
                                                  Prob > chi2       =     0.0002
Log likelihood =  -79.34221                       Pseudo R2         =     0.1315
```

attack	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
smokes	1.544053	.3998329	3.86	0.000	.7603945	2.327711
age	.026112	.017042	1.53	0.125	-.0072898	.0595137
bmi	.1129938	.0500061	2.26	0.024	.0149837	.211004
hsgrad	.4048251	.4446019	0.91	0.363	-.4665786	1.276229
female	.2255301	.4527558	0.50	0.618	-.6618549	1.112915
_cons	-5.408398	1.810603	-2.99	0.003	-8.957115	-1.85968

ageは有意ではない

BMIの欠損値

. misstable sum

Variable	Obs=.	Obs>.	Obs<.	Obs<. Unique values	Min	Max
bmi	22		132	132	17.22643	38.24214

- 変数の中でBMIにだけ欠損値が22個存在する
- 推定ではこの22個行のデータがリストワイズに除去されている

MIのフロー

mi set:データの形式(並び方)の設定

↓

mi register:代入したい変数(欠損値のある変数)
mi regular:欠損値のない変数

↓

mi impute:欠損値への代入方法の設定

↓

mi estimate:モデル(パラメータ)の推定

MIのデータ形式

. mi set mlong

- mi setはデータセットの並び方を設定する。データの保存形式にはflongsep, flong, mlong, wideの4種類がある
- flongsepは代入したデータセットを個別に作成する。例えば、pat.dtaというファイルであれば、_1_pat.dta, _2_pat.dtaという名前のdtaファイルを作成する
- flongは縦に代入したデータセットを追加する。データセットは最終的に_N=N+M×N行になる
- mlongも縦に並べるが、_N=N+M×n行となる。nは代入したデータの行数
- wideは横方向に並べていくので、行数は_N=Nで変化しない

変数の登録

```
. mi register imputed bmi
. mi register regular attack smokes age hsgrad female
```

- 欠損値のある変数(代入したい変数)をmi register imputedで指定する
- 欠損値のない変数はmi register regularで設定する
- ファイルを保存したら次回以降、この作業は不要
- 設定内容を変更するときは、再度、新しい情報でコマンドを実行する(上書き)
- 設定情報をクリアするときは、mi set,clear

代入モデル

```
. mi impute regress bmi attack smokes age hsgrad female,
      add(20) rseed(2232)

Univariate imputation                          Imputations =       20
Linear regression                                    added =       20
Imputed: m=1 through m=20                          updated =        0

------------------------------------------------------------------------
                |                Observations per m                |
                |------------------------------------------------------|
       Variable |   Complete   Incomplete    Imputed |      Total
----------------+-------------------------------------+----------------
            bmi |        572           22         22 |          594
------------------------------------------------------------------------
(complete + incomplete = total; imputed is the minimum across m
 of the number of filled-in observations.)
```

代入したデータを20セット作成

解析モデルの被説明変数を必ず利用する

22個の欠損値に代入実行

欠損のない完全な行が572行存在する

代入データの確認

. mi xeq 0 1 20:summarize bmi

- mi xeqは代入した個別のシート(メモリー上にある20枚)に同じコマンドを繰り返し適用する

```
m=0 data:
-> summarize bmi

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         bmi |        572    25.05037    4.173393   14.48172   38.24214

m=1 data:
-> summarize bmi

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         bmi |        594    25.07279    4.152333   14.48172   38.24214

m=20 data:
-> summarize bmi

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         bmi |        594     25.149    4.187616   14.48172   38.24214
```

モデル推定と統合

. mi est, dots:logit attack smokes age bmi hsgrad female

attack	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
smokes	.6227489	.2175802	2.86	0.004	.1962994	1.049198
age	.0853274	.0099554	8.57	0.000	.0658151	.1048397
bmi	.0954405	.0252816	3.78	0.000	.0458867	.1449943
hsgrad	-.2932975	.2320512	-1.26	0.206	-.7481095	.1615145
female	-2.081514	.3034205	-6.86	0.000	-2.676208	-1.486821
_cons	-7.65917	.9618609	-7.96	0.000	-9.544415	-5.773926

- MIを利用した結果、ageが有意になった。つまり、年齢が高くなると、心臓発作のリスクが高くなる事が分かる
- 20セットの完全データから得た推定値を統合した

ポイント

- MIはシミュレーションを利用したモデル推定の手法。欠損値そのものを復元するものではない
- 1)欠損データに対して適切な代入方法を利用した、2)欠損値が存在しないと考えた時のモデルが適切である、という2つの仮定が正しい場合に、 MIを利用した推定結果の解釈は正当性を持つ
- 欠損の程度によって代入の回数は決める。欠損が少ない場合は5-20回程度。欠損の程度が大きくなるほど、回数を増やす。
- 代入回数が少ない場合、パラメータの検定統計量はt分布(F分布)を利用する。残差の自由度は回数Mに依存する
- 逆に代入回数が多い場合、正規分布(カイ二乗分布)を利用する
- 解析モデルと代入モデルのそれぞれについて感度分析を行うことが望ましい
- 尤度や逸脱度などの統計量を普通の統計解析と同じ視点で解釈することはできない

多重代入の基礎知識

- 欠損値に関する仮定
- Missing completely at random(MCAR)…欠損する確率がどの変数にも依存していない

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

  - 例えば、血圧がYで年齢がXとする。転勤の辞令で引っ越しした ので、定期的な追跡検査から脱落し、データが欠損する(転勤の辞令が個人の変数とは無相関とする)
  - この時、欠損値はランダムに発生すると考えられ、先のロジットモデルにおけるリストワイズなデータの除外による推定結果は正当性を持つ(バイアスなし)

MAR

- Missing at random(MAR)…欠損する確率がある変数に依存する

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

  - 同じく、血圧がYで年齢がXとする。ある年齢に達すると転勤するような場合、欠損する確率は年齢に依存すると考える
  - MARの場合にリストワイズな削除を行うと、推定値にバイアスが生じる
- Missing not at random(MNAR)…あまりにも血圧Yが高すぎて、検査から脱落
  - MNARはモデルを変更して対応する

Ignorable

- 欠損値の発生確率のメカニズムがMARで、モデルのパラメータと欠損データのメカニズムに関するパラメータが異なるとき、欠損データのメカニズムはignorable(無視できる)
- ignorableの時、分析者は代入モデルだけ考えれば良い
- MNARだとignorableにならない
- ignorabilityの仮定は検定ではチェックできない(欠損値は復元できない)ので、合理的な判断が必要
- 感度分析を利用してモデルの確かさを確認する

mi impute regress

- 代入方法の中ではポピュラーな方法の一つ
  - 欠損値のある変数の正規性を仮定していることに注意
  - できるだけ正規性を保つために、変換することも考える
  - 例題のbmiに欠損値が存在する。これをxとし、回帰モデルのattack smokes age hsgrad femaleをzとする

$$x_i | z_i \sim N(z_i' \beta, \sigma^2) \qquad (1)$$

  - 係数βと分散σ2が回帰モデルのパラメータ
  - データを観測できているグループ0(ゼロ)と欠損しているグループmに分ける

$$x = (x_0, x_m), \; z = (z_0, z_m)$$

mi impute regress

$$x_i|z_i \sim N(z_i'\beta, \sigma^2) \qquad (1)$$

Step1.1式を観測値グループ0で推定してβとσ2の推定値を得る

Step2.回帰モデルの誤差分散を自由度(n0-q)のカイ二乗値の乱数で割って、新たな誤差分散を発生させる。その誤差分散と、回帰モデルの係数推定値による正規分布の乱数から、代入モデルの推定値βを得る。qは定数項を含む回帰モデルの説明変数の数

$$\sigma_*^2 \sim \hat{\sigma}^2(n_0 - q)/\chi^2_{n_0-q}$$

$$\beta|\sigma_*^2 \sim N\left\{\hat{\beta}, \sigma_*^2(Z_0'Z_0)^{-1}\right\}$$

mi impute regress

Step3.Step2で得たパラメータを使ってxの代入値を得る
Step4.Step2とStep3を繰り返してMセットの完全データセットを得る

Step2とStep3はposterior predictive distribution(事後予測分布)を利用したシミューレション(Gelman et al.[2014, 354-357])と呼ばれている

欠損パターン

- 直感的に分かりやすい線形回帰モデルを代入モデルとする方法で全部解決できるか?
- この方法は欠損のある変数が連続変数で正規分布に従うという仮定のもとに利用する
- 正規性の仮定が成立しない場合は利用しない

まとめ

- 利用頻度の高いmi impute regressを利用してMIの実行例を紹介した
- Stataで多重代入を利用する場合の手順を理解しよう
- 多重代入に関する重要なキーワードを説明した。欠損の仕組みがMARで、ignorableの仮定が成り立つとき、代入モデルを利用して多重代入が利用可能となる
- 代入モデルには解析モデルの被説明変数を必ず利用する
- 次回は欠損値への一般的な対応方法について、いくつかの手法を比較しながら詳しく説明する

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。

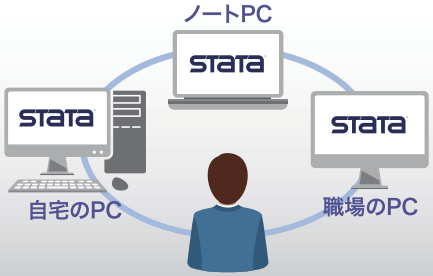
ライセンスはサブスクリプションがおすすめ！

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

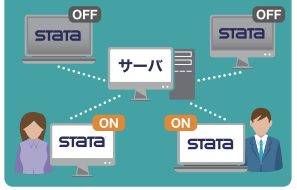
シングルライセンスを複数人でまとめてご購入いただく際のボリュームディスカウント商品です。

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



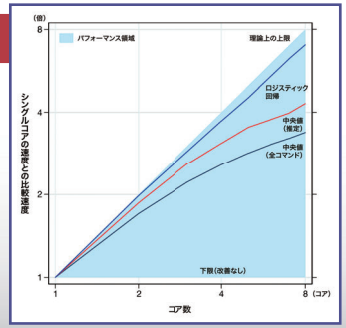
1台の仮想マシンにユーザがアクセスする運用
例:2ライセンスの場合

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/