第 1 章のデータとプログラムについての説明

慶應義塾大学　星野崇宏

ここでは

1）図 1.3 のプログラム

2）例 1.3 のプログラム

を紹介する。

例 1.2 と例 1.4 は 6 章を参照されたい。

【1 ：図 1.3 のプログラム】

「従属変数に説明変数の欠測確率が依存する」場合のうち、一番単純な例として図 1.3 に記載した「従属変数が一定値 C 以下の対象は説明変数が欠測する」という状況のシミュレーションを行う。さらに、平均値代入を行うとバイアスが増大することも示す。モデルは

$$y=\alpha+\beta x+e,\quad \text{真値}\,\alpha=0.2,\quad \beta=0.3,\quad V(e)=40^2,\quad E(x)=600,\quad V(x)=100^2$$

＊結果として相関係数の真値 0.6

＊x が世帯収入,y が家賃などを想定するとよい。収入が低い人は収入への回答を行わないので、完全ケースだけから解析をすると所得の家賃への影響度が低く推定されるような場合。

プログラムは regCCmissingx.r である。以下に説明する。

cmatr2 <- cmatr[order(cmatr$y, decreasing=T),]

は平均値代入のために、後半の人たちに説明変数の欠測が起きるようなデータにしている。

omatr <- subset(cmatr2, cmatr2$y>C)

で完全ケースを作成する。

ylm1 <- lm(cmatr2$y ~ cmatr2$x)

は完全データに回帰分析を実施したもので、サンプルサイズ N を大きくすると当然ながら回帰係数は正しく推定される。一方

ylm2 <- lm(omatr$y ~ omatr$x)

は完全ケースへの回帰分析で、回帰係数の真値が 0.3 だが推定値 0.2 程度であり、これは N を大きくしてもバイアスが残る。

一方、x と y の役割を変える“逆回帰”を行うと、

xlm2 <- lm(omatr$x ~ omatr$y)

完全ケース解析であっても回帰係数は以下の完全データでの解析

xlm1 <- lm(cmatr2$x ~ cmatr2$y)

の結果にほぼ一致し、N が大であれば真値に収束する一致性を有する。

＊ちなみに完全データにおいて通常の回帰と逆回帰での回帰係数は当然ながら一般的に一致しないことに注意されたい。

このことは、全てのオブザベーションに対して観測されている y にのみ欠測確率が依存している場合には、条件付き分布 p(y|x)の母数の推定では「ランダムでない欠測」（1 章図 1.7 の④）であるが、条件付き分布 p(x|y)の母数の推定においては「ランダムな欠測」（1 章図 1.7 の③）になること、そしてこの場合に直接尤度の最大化は完全ケースの最小二乗基準の最小化と同じことを行っていることから、逆回帰をして得た係数をもとの(x で y を説明する)回帰係数に戻す操作をすればこの場合の正しい回帰係数を出すことができる。詳しくは 2 章や 5 章を参照されたい。

さらに、説明変数の欠測値部分に対して、観測値を用いて平均値代入した場合もプログラムには記載されている。データセットは y の値でソーティングされているので、後ろの方の（y の値が低い）オブザベーションの x について平均値代入し、得られた“疑似完全データ”mimpd を使って回帰分析や相関係数の推定を行うと、先ほどの完全ケース解析よりもさらにバイアスはひどくなる。

平均値代入は実務では何となく利用してしまいがちである。上記のような 2 変数でこのようなことは行わないかもしれないが、複数変数がある場合に何も考えるとお手軽な方法として実施すると大きなバイアスを生じることに注意されたい。

【２：例 1.3 のプログラム】

データの詳細は本文中を参照されたい。ここでは NLSY データのうち、R のライブラリの library(mi)に掲載されているデータセット nlsyV を利用した解析のプログラムを説明する。

library(mi)はもともとコロンビア大学の Gelman 教授らのグループが開発したものであり、連鎖式による多重代入（4 章 7 節参照）を行うだけではなく、欠測パターンごとの記述統計の提示やパターンの図示を行うことができる。

プログラムは Minlsy.r である。以下に重要な部分を説明すると

表 1.2 の完全ケースの解析は説明変数の一部を factor 型などに適切に変換後

```
nlsyVlm <- lm(ppvtr.36 ~ first + b.marr + income + momage + momed + momrace, data = nlsyV2)
```

で実施するが、こちらでは 228 人中すべての変数に欠測がない 172 人のデータからの解析となっている。

#####以下結果#####

```
Call:
lm(formula = ppvtr.36 ~ first + b.marr + income + momage + momed +
    momrace, data = nlsyV2)

Residuals:
     Min       1Q   Median       3Q      Max
-37.050 -10.762   1.454  10.421  34.172
#nlsyV は本文にも記載した national longitudinal study of youth データの部分データ
#下記は表 1.2 の完全ケースでの回帰分析の結果
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.007e+01  1.162e+01   6.029 1.08e-08 ***
first        5.266e+00  2.717e+00   1.938 0.054375 .
b.marr       6.198e+00  3.372e+00   1.838 0.067894 .
income       6.100e-06  1.222e-05   0.499 0.618154
momage       1.778e-01  4.490e-01   0.396 0.692567
momed.L      6.164e+00  4.421e+00   1.394 0.165176
momed.Q     -3.151e+00  3.282e+00  -0.960 0.338527
momed.C     -1.895e+00  2.492e+00  -0.761 0.447955
momrace2    -3.890e+00  4.195e+00  -0.927 0.355120
momrace3     1.415e+01  3.894e+00   3.634 0.000375 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.73 on 162 degrees of freedom
  (228 observations deleted due to missingness)
Multiple R-squared:  0.3722,    Adjusted R-squared:  0.3373
F-statistic: 10.67 on 9 and 162 DF,  p-value: 6.471e-13
########解析結果終わり#####
```

missing_data.frame()は library(mi)の提供するデータフレーム型に変換する関数であり、実際には個別の変数について順序付きカテゴリーか、名義カテゴリーか、連続値かなどを指定することで、背後で完全付き条件分布の指定が自動的に行われる。

mi(データフレーム, n.iter = , n.chains = , max.minutes =)
は代入ステージの関数であり。n.iter は代入の繰り返し数の上限。n.chain は MCMC の並

列なチェーンとこの場合同義である。他の引数は help を参照されたい。

上記の関数から得られた多重代入データに対して
pool(glm の式,data=多重代入データ)
という関数を適用することで、解析フェーズと統合フェーズを実施してくれる。

#####以下結果#####
#下記は表 1.2 の多重代入での回帰分析の結果

	coef.est	coef.se
(Intercept)	78.94	8.19
first1	4.14	1.88
b.marr1	4.64	2.13
income	0.00	0.00
momage	0.00	0.35
momed.L	11.20	2.98
momed.Q	1.07	2.28
momed.C	0.22	2.00
momrace2	-5.65	3.77
momrace3	11.82	3.14

n = 390, k = 10
residual deviance = 92118.7, null deviance = 140425.8 (difference = 48307.2)
overdispersion parameter = 236.2
residual sd is sqrt(overdispersion) = 15.37

#人種だけでなく、第一子（少子と親の投資の内生性)、子供の誕生時に結婚している、大卒で高い。完全ケースより情報量が多いことを反映して、明らかに多くの係数が有意になっている。
但し連鎖式による多重代入の問題は 4 章 7 節を参照されたい。
##########結果終わり##########