傾向スコアマッチング：実行のためのメモ

渡邊直樹

2021 年 10 月 27 日

EZR での手続き

ここには，EZR（Easy R）で傾向スコアマッチングを行う手順をまとめておく．傾向スコアマッチングでは 2 項ロジスティック回帰を援用する．2 項ロジスティック回帰の解説については，別の資料を参照してほしい．また，自分の計算機にデータセット（data3.csv）が保存されているとして，それを EZR で読み込むことにする．R（EZR）は 2 バイト文字を読み込めないことに注意せよ．データが格納されているフォルダが日本語で書かれていると，データを読み込むことができない．計算機のデスクトップにデータを置いておくとよいだろう．

R（EZR）で傾向スコアマッチングを行うには，Matching パッケージを CRAN のミラーサイトからダウンロードして，インストールする必要がある．R の特徴は最新のパッケージを必要に応じてダウンロードして使えることにあり，高価な統計ソフトウェアにはまだ組み込まれていないデータ解析手法でもこの機能により実行可能になる．本日の統計パートでの実習はパッケージをインストールして EZR で使う練習を兼ねている．

1. データの読み込み：リボンに表示されているボタンを次の順で押す．
    - データ＞データのインポート＞テキストファイルまたはクリップボード，URL から
    - ポップアップウィンドウにおいて，フィールドの区切り記号として「カンマ [,]」を選択し，OK ボタンを押す．デフォルトでは「空白」が選択されている．（データファイルの場所はデフォルトでローカルファイルシステムが選択されているので，そのままにしておく．）

- 新たなポップアプウィンドウが現れるので，そこで data3.csv を選択し、Open ボタンを押す.
- ホーム画面上で「データセットを表示」ボタンを押して，データが読み込まれていることを確認する.
- 時々、EZR がデータセットを読み込まないことがある．その場合には，エクセルファイル（xls または xlsx）で保存し直して，データ＞データのインポート＞ Excel ファイルから... を押して読み込むと大抵はうまくいく.

2. 2 項ロジスティック回帰：リボンに表示されているボタンを次の順で押す.
   - 統計量＞モデルへの適合＞一般化線型モデル
   - ポップアップウィンドウの「データ」の欄にて，treatment をダブルクリックして選択し，被説明変数とする．次に，treatment 以外のすべての変数を順次ダブルクリックして説明変数とする.
   - リンク関数族が binomial に，リンク関数が logit になっていることを確認して,「適用」または「OK」ボタンを押す.
3. 傾向スコアを保存 : リボンに表示されているボタンを次の順で押す.
   - モデル＞計算結果を保存...
   - ポップアップウィンドウで「予測値」の項目にのみチェックを入れて,「OK」ボタンを押す.
   - 以上の操作は EZR の R スクリプトの欄に直接
     Dataset$ PS=GLM.1$fitted.values
     と打ち込んで「実行」ボタンを押してもよい.

EZR の出力は次のとおりである.（とりあえず，以外の変数をすべて説明変数として回帰式に組み込んだ.）

```
Call:
glm(formula = treatment ~ age + BMI + day + EF + GES + Hgb +
    PVC + sex + UCr, family = binomial(logit), data = Dataset)
```

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.82846  -0.41755  -0.06904   0.38785   2.22255

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.985898  12.460119   0.561  0.57503
age          0.262986   0.094534   2.782  0.00540 **
BMI         -0.469486   0.409614  -1.146  0.25173
day         -0.571267   0.206020  -2.773  0.00556 **
EF          -0.117527   0.082207  -1.430  0.15282
GES          0.020521   0.009784   2.097  0.03595 *
Hgb         -0.526926   0.496367  -1.062  0.28843
PVC         -0.018030   0.057337  -0.314  0.75317
sex          0.504913   0.928511   0.544  0.58659
UCr         -2.526849   2.522135  -1.002  0.31641
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 76.382  on 59  degrees of freedom
Residual deviance: 36.974  on 50  degrees of freedom
AIC: 56.974

Number of Fisher Scoring iterations: 7


> exp(coef(GLM.1))  # Exponentiated coefficients ('odds ratios')
  (Intercept)       age           BMI           day           EF
1081.27746633   1.30080796   0.62532348   0.56480956   0.88911675
   GES           Hgb           PVC           sex           UCr
1.02073349   0.59041696   0.98213138   1.65684059   0.07991044

> Dataset <- within(Dataset, {
```

```
+     fitted.GLM.1 <- fitted(GLM.1)
+ })
```

EZR のリボンにある「データセットを表示」ボタンを押して，傾向スコア（2 項ロジスティック回帰の予測値）が書き込まれているかを確認する．変数名は fitted.GLM.1 となっている．

4. 傾向スコアマッチング：Matching パッケージを使う．

   - まず，treatment のダミー変数を作成する．Windows 版では，アクティブデータセット＞変数の操作＞ダミー変数を作成するというボタンを押す．ポップアップで treatment を選択し，ダミー変数であることを示す文字列として，Dummytreatment と入力し，OK ボタンを押す．）
   - Mac 版にはこれらのボタンがない．よって，ここでは Excel で data3.csv を操作して，次のようなダミー変数を作成する．treatmentDummytreatmentA では新薬の投与を受けた者を 1，投与を受けていない者を 0 と入力する．tratmentDummytreatmentB では新薬の投与を受けた者を 0，投与を受けていない者を 1 と入力する．（予めこの作業を完了した後で，データを読み込むとよい．データを読み込んだ後でデータを編集したい場合には，EZR のリボンに「データセットの編集」というボタンがあるので，それを使ってもよい．）
   - EZR の R スクリプト画面で次のようにタイプし，（カーソルがその行の末尾にある状態で）「実行」ボタンを押す．
     install.packages(”Matching”)
     - しばらくすると，secure CRAN mirrors というポップアップが出てくるので，適当なサイト（たとえば Japan (tokyo) [https]）をクリックして，「Ok」ボタンを押すと Matching パッケージのダウンロードが開始される．出力画面に The downloaded binary packages are in... といった文字列が出てきたら，ダウンロードに成功したと判る．
   - R スクリプト画面で次のようにタイプし，カーソルがその行の末尾にある状態で）「実行」ボタンを押す．これで，atching パッケージが R に読み込まれる．
     library(Matching)

– 次のような文字列がメッセージ画面に出てきたら，パッケージの読み込みにに成功したと判る．

```
## Jasjeet S. Sekhon. 2011. ‘‘Multivariate ...
## Software with Automated Balance ...
## Journal of Statistical Software, 42(7): 1-52.
##
```

- R スクリプト画面で次のように「1 行で」タイプし，カーソルがその行の末尾にある状態で）「実行」ボタンを押す．
  Matching <- Match(Y=Dataset$day,
  Tr=Dataset$treatmentDummytreatmentA,
  X=Dataset$fitted.GLM.1, caliper=0.25, ties=F, replace=F)
  – 出力画面にもメッセージ画面にも，何も表示されなかったら，マッチングがうまくなされたことを意味する．
- R スクリプト画面で次のようにタイプし，（カーソルがその行の末尾にある状態で）「実行」ボタンを押す．
  summary(Matching)

  出力画面には次のような結果が表示される．

```
Estimate...  1
SE.........  1.4252
T-stat.....  0.70165
p.val......  0.4829

Original number of observations..............  60
Original number of treated obs...............  20
Matched number of observations...............  8
Matched number of observations  (unweighted).  8

Caliper (SDs)..............................   0.25
Number of obs dropped by 'exact' or 'caliper'  12
```

  60 人分のサンプルでマッチされたのは 8 組のみ．これでは 2 群の平均を比較しても，大半のデータを捨ててしまっていることになる．そこで，傾向スコアの大きさに基づいて，いくつかの層（サブクラス）に分け，次のような計算を行うことで，全て

のデータを使うこともある．実務では5つの層に分けることが多いので，ここでもそのようにしておこう．全体での調査対象者の数を n，処置群と対照群に属する第 k 層の調査対象者の数を n_k，処置群に属する第 k 層の調査対象者のデータの平均を $\bar{y}_1^k$，対照群に属する第 k 層の調査対象者のデータの平均を $\bar{y}_0^k$ とすると，因果効果 $E(y_1 - y_0)$ は

$$\sum_{k=1}^{5} \frac{n_k}{n}(\bar{y}_1^k - \bar{y}_0^k)$$

と推定される．この方法と傾向スコアの近いサンプルをマッチさせる方法のどちらが良いかは一概にはいえない．

ここで，使った順番にコマンドをリストしておく．

- install.packages("Matching")
- library(Matching)
- Matching <- Match(Y=Dataset$day, Tr=Dataset$treatmentDummytreatmentA, X=Dataset$fitted.GLM.1, caliper=0.25, ties=F, replace=F)
- summary(Matching)

5. ペアマッチされたデータを抽出する．スクリプト画面で次のようにタイプし，カーソルがその行の末尾にある状態で）「実行」ボタンを押す

    - スクリプト画面で次のようにタイプし，カーソルがその行の末尾にある状態で）「実行」ボタンを押す．

    ```
    treat <- Dataset[Matching$index.treated,]
    control <- Dataset[Matching$index.control,]
    Match.data <- rbind(treat, control)
    ```

以上のコマンドにより，新たなデータセット Match.data が生成される．リボンの下にある「Dataset」をクリックして，出てきたポップアップで　Match。data を選択すると．データを表示できる．保存するには次のようにする．

- データ＞アクティブデータセット＞アクティブデータセットのエクスポート.（アクティヴデータセットの保存を押すと，RData 形式で保存されてしまうので，Excel などで読み込めない.）ポップアップウィンドウにおいて，フィールドの区切り記号として「カンマ [,]」を選択し，OK ボタンを押す．デフォルトでは「空白」が選択されている．

スクリプト画面においてコマンドを入力する際には，大文字か小文字か，ピリオドがあるかないか，に気をつけること．うまくいかない時は，大抵，こうした入力ミスが原因である．