差分の差分法と欠測データの補完

渡邊直樹

2023 年 11 月 4 日

差分の差分法（Difference-in-Differences method）分とは，ある処置を施されたグループ（処置群）とそうではないグループ（対照群）の間で時間の経過を伴う結果を比較することで，処置の因果効果をより正確に推定するための統計学的手法のことである．本稿では，韓国租税財政研究所（Korea Institute of Taxation and Finance）の NaSTaB に納められているデータのうち，2013 年度から 2018 年度に調査されたものを使用し，韓国における大統領の弾劾前後での政治的傾向の変化が慈善的寄付金総額にどのように影響したかを検証する．

1　分析手順の概略

NaSTab では回答者に自身の政治的傾向（保守的，中立的，進歩的）を回答させているが，これまでに勉強した 2 項ロジスティック回帰やそれに基づく傾向スコアマッチングを利用しやすくするため，データセットでは，大統領の弾劾前後で中立的な立場から進歩的な立場に変化させたかどうかだけで処置群と対照群に分ている．韓国では政治的に進歩的立場にある個人の方が慈善的寄付金に応募する人が多いようなので，単純化のため，保守的立場をとる個人は慈善的寄付金には応募しないと仮定した．

各回答者に付与された ID を i，データ取得年度を t で表すと，回答者 i の政治的傾向の変化が寄付金額の変動に及ぼす影響を次の差分の差分法（DID）を表現する線形回帰モデル (1) で捉えてみよう．被説明変数 lnGiving は年間寄付支出総額の自然対数を取った変数である．

$$\begin{aligned} \mathrm{lnGiving}_{it} &= \beta_0 + \beta_1 \mathrm{treatment}_i + \beta_2 \mathrm{after} + \beta_3 \mathrm{treatment} * \mathrm{after} \quad (1) \\ &\quad + \beta_4 x_{it}^1 + \beta_5 x_{it}^2 + \cdots + \epsilon_{it} \end{aligned}$$

ここで，$x_{it}^1, x_{it}^2, \ldots$ は回答者 i の t 年度における社会経済的特徴を表す変数を略記したものであり，変数 educationl_{it} $(l = 1, 2, \ldots, 6)$，indepm_{it}（$m = 1, 2, \ldots, 5$），region$_{it}$ を含む．各変数の説明は「データソース：National Survey of Tax and Benefit」（データソース.xlsx）を参照してほしい．なお、x_{it}^1 は年間総所得の自然対数をとった値をとる変数（indep1）なので，係数 β_4 は回答者の年間総所得が 1%増加したときに年間寄付金支出総額が β_4%増加することを意味する．同様に，x_{it}^2 は回答者が所有する住宅の価格の自然対数をとった値をとる変数（indep2）なので，β_5 もそのような「弾力性」を意味する係数である．

変数 treatment は処置群（treatment group）か対照群（control group）かを表すダミー変数であり，大統領の弾劾後に中立的な政治的立場を進歩的なものに変化させた場合は 1 を，中立的な政治的立場を変化させなかった場合には 0 を取る．変数 after は大統領の弾劾前後を表すダミー変数であり，それがなされた 2016 年以降の 3 年度は 1，それ以前の 3 年度では 0 を取る．DID を表現するモデルとして，線形回帰モデル (1) は最も単純で標準的なものではあるが，β_1，β_2，$\beta 3$ が何を計測しているかを明確にしておこう．以下では，$E(x|y = 0, z = 1)$ は条件 $y = 0$ かつ $z = 1$ の下での確率変数 x の期待値を表す．期待値という用語が分かりにくければ,（正確な表現ではないが）平均値と読み替えて，統計的操作の概略を理解してほしい．

- after $= 0$ のとき：

$$\begin{aligned} &E[\text{lnGiving}|\text{treatment} = 1, \text{after} = 0] \\ &-E[\text{lnGiving}|\text{treatment} = 0, \text{after} = 0] = \beta_1. \end{aligned} \tag{2}$$

- treatment $= 0$ のとき：

$$\begin{aligned} &E[\text{lnGiving}|\text{treatment} = 0, \text{after} = 1] \\ &-E[\text{lnGiving}|\text{treatment} = 0, \text{after} = 0] = \beta_2. \end{aligned} \tag{3}$$

- treatment $= 1$ のとき：

$$\begin{aligned} &E[\text{lnGiving}|\text{treatment} = 1, \text{after} = 1] \\ &-E[\text{lnGiving}|\text{treatment} = 1, \text{after} = 0] = \beta_2 + \beta_3. \end{aligned} \tag{4}$$

(2)，(3)，(4) より，β_3 は「処置群での変化」と「対照群での変化」の差分，つまり,「差分の差分」で表される．

$$
\begin{aligned}
\beta_3 \quad = \quad & (E[\text{lnGiving}|\text{treatment} = 1, \text{after} = 1] \\
& \qquad -E[\text{lnGiving}|\text{treatment} = 1, \text{after} = 0]) \\
& -(E[\text{lnGiving}|\text{treatment} = 0, \text{after} = 1] \\
& \qquad -E[\text{lnGiving}|\text{treatment} = 0, \text{after} = 0])
\end{aligned} \tag{5}
$$

以上より，係数 β_3 は処置の非説明変数に対する因果効果となっている．したがって，線形回帰モデル (1) の係数 β_3 の推定値は回答者が政治的立場を変化させたことによる慈善的寄付金への効果を計測していることになる．

1.1 欠測（treatment）を含むデータを全て削除した場合

元のデータセット（practice）から treatment に欠測があるデータを全て削除したデータセット（practice2）を作成した．Excel でデータ＞並べ替えを使った.（データの順番を変えたくなければデータ＞フィルターを使うとよい.）EZR で practice.xlsx を読み込んで，アクティヴデータセット＞「欠損値のあるケースを削除」などで操作してもよい．

- カーソルで treatment の行を選択し,「並べ替え」ボタンを押す．
- 「並べ替えの前に」というポップアップが出てくるので，選択範囲を拡張する」にチェックを入れ,「並べ替え...」ボタンを押す．
- 「並べ替え」というポップアップが現れるので,「最優先させるキー」をクリックし，treatment を選択する.「順序」は最小から最大でも，最大から最小でもよい．OK ボタンを押す．

セルに 0 か 1 の値が入っているデータは 7046 件ある.（変数名が入っている 1 行目を含まない.）それ以外のデータを全て削除する．これくらいのサンプルサイズでも Excel では回帰分析を実行できないことが多い．次に，DID 回帰式で必要となる変数 $treatment * after$ を作成する．

- 変数 after と region の間に一行挿入し，変数名 $treatment * after$ を記入する.（R 列）

- R2 セルに=P2*Q2 と入力して，return キーを押す．
- R2 セルをコピーし，最終行までペーストする．R2 セルの右端にカーソルを持っていくと + マークが出てくるので，ダブルクリックする．

統計解析＞連続変数の解析＞線形回帰（単回帰、重回帰）の順にクリックし，被説明変数は lnGiving，説明変数は以下の call:における after 以降のものを選択する．

```
Call:
lm(formula = InGiving ~ after + after.treatment + education1 +
    education2 + education3 + education4 + education5 + education6 +
    indep1 + indep2 + indep3 + indep4 + indep5 + region + treatment,
    data = Dataset)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.747383   0.210729  -8.292  < 2e-16 ***
after            0.073024   0.057644   1.267  0.20526
after.treatment  0.165140   0.090580   1.823  0.06832 .
education1       0.143904   0.117452   1.225  0.22054
education2       0.179662   0.097782   1.837  0.06620 .
education3       0.565886   0.113496   4.986 6.32e-07 ***
education4       0.712810   0.104277   6.836 8.85e-12 ***
education5       1.672552   0.154518  10.824  < 2e-16 ***
education6       1.918888   0.243871   7.868 4.14e-15 ***
indep1           0.296051   0.029278  10.112  < 2e-16 ***
indep2           0.014120   0.004907   2.877  0.00402 **
indep3          -0.003428   0.023607  -0.145  0.88456
indep4          -0.654963   0.071123  -9.209  < 2e-16 ***
indep5           0.064143   0.075609   0.848  0.39627
region           0.012314   0.002604   4.728 2.31e-06 ***
treatment       -0.031290   0.065647  -0.477  0.63363
```

係数 $treatment * after$ の p 値が 0.06832 となっており，微妙な結果．

さらに 2015 年と 2016 年以外のデータを全て削除した場合

韓国で元大統領の弾劾裁判があったのは 2015 年から 2016 年にかけてのことなので，year=2015 または 2016 以外のデータを削除する.（上述の方法を参照せよ.）

```
Call:
lm(formula = InGiving ~ after + after.treatment + education1 +
    education2 + education3 + education4 + education5 + education6 +
    indep1 + indep2 + indep3 + indep4 + indep5 + region + treatment,
    data = Dataset)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.411391   0.363358  -3.884 0.000105 ***
after           -0.004626   0.099553  -0.046 0.962944
after.treatment  0.308441   0.156088   1.976 0.048260 *
education1       0.171803   0.203074   0.846 0.397632
education2       0.223742   0.168716   1.326 0.184917
education3       0.645657   0.195767   3.298 0.000988 ***
education4       0.739518   0.179158   4.128 3.79e-05 ***
education5       1.641685   0.262983   6.243 5.07e-10 ***
education6       2.007105   0.418708   4.794 1.74e-06 ***
indep1           0.284432   0.049956   5.694 1.39e-08 ***
indep2           0.013152   0.008508   1.546 0.122289
indep3          -0.014400   0.040829  -0.353 0.724348
indep4          -0.676368   0.120878  -5.595 2.45e-08 ***
indep5           0.089987   0.131130   0.686 0.492626
region           0.007203   0.004490   1.604 0.108795
treatment       -0.154450   0.112174  -1.377 0.168676
```

流石に，クリアな結果が出た.

2　欠測データの補完

本実習で用いるデータセットの元になった調査（NaSTab）では，自らの政治的傾向を表明しなかった回答者が多数存在した．そのため，政治的傾向に変化があったかどうかに関する変数 treatment にも多くの欠損値が見られる．ここでは，2 項ロジスティック回帰を行なって傾向スコアを求め，各回答者の傾向スコアが同じくらいのデータで欠測値を補完する方法を紹介する．

1. 表明された政治的傾向に関するものではなく，回答者が政治的傾向に対する質問に回答したか否かに関する 2 項ロジスティック回帰を適用し，その予測値を傾向スコアとする．

2. 傾向スコアの大きさによって回答者をいくつかの層に分ける．(5 つの層に分けることが多い．)

3. 2. で区分けした各層 k（$k =, 2, \ldots, 5$）において，政治的傾向を表明しなかった回答者からなるグループを N_0，それを表明した回答者からなるグループを N_1 とする．グループ N_1 から $|N_0|$ 個のデータをランダムに復元抽出して，政治的傾向を表明していない回答者の潜在的な政治的傾向を表す値として用いる．

上記の補完を行えば，すべての年度における全回答者について，政治的選好（polipref）の欠測値が補完されたデータセットを作成することができる．しかし，エクセル統計ではサンプルサイズが多く過ぎて計算にものすごく時間がかかるか，途中で計算が止まってしまう．以下では，前章と同様に，polipref ではなく treatment の欠損値に注目する．

まず，傾向スコアを算出する．

- 変数 treatment の左隣に列を挿入し，P 列とする．変数名を，たとえば，answered としておく．(P1 セルに記入する．)

- P2 セルに=if(Q2=””, 0, 1) と入力し，return キーを押す．(P2=が空白であれば，P2 セルに 0 を代入し，そうでなければ 1 を代入せよ，という意味．文字列は” ”で囲む．”は shift+ 2 のキーボードに対応する記号．)

- P2 セルをコピーし，最終行までペーストする．R2 セルの右端にカーソルを持っていくと + マークが出てくるので，ダブルクリックする．

- エクセルファイルを一旦保存し，EZR で読み込む．ここではファイル名を practice3 としておく．
- 被説明変数を answered，説明変数を education1, ... education 6, indep1, ... , indep5，region とする 2 項ロジスティック回帰を行う．
  - 名義変数を因子に変換しておく．変数 answered だけでなく，indep4 は性別，indep5 は marital status なので，これらも名義変数である．教育水準（education1, ..., education6 ）も同様だが，あまり問題とはならないので，因子に変換してもしなくてもよい．
  - 統計解析＞名義変数の解析＞二値変数に対する多変量解析（ロジスティック回帰）の順にクリック.
  - EZR では「傾向スコアを自動作成する」オプションにチェックを入れておくこと．
  - 政治的選好 polipref を説明変数に組み込むと，欠損値となっているものについては傾向スコアが計算されない．
  - 年間寄付総額 lnGiving は，DID 回帰式の推計では被説明変数となるため，傾向スコアを計算する際には説明変数にしない方がよいだろう．
- データを傾向スコアで 5 等級に区分する．
  - 標準メニュー＞データ＞アクティブデータセット内の変数の管理＞数値変数を区間で区分＞ の順にクリック.
  - スライドバーで区分の数を 5 つにする.（ランダムサンプリングせずに，同じ区分内の平均を参照して欠測値を補完するならば，区分の数はもう少し細かくてもよい.）
  - 区分の方法は k-平均クラスタリングでもよいが，クラスタリングについては別の講義資料を参照する必要があるので，ここでは「等間隔」を選択しておく．
- 5 等級の区分がデータセットの末尾に加わっていることを確認する.（「データセットの編集」ボタンを押すと，データセットの中身を見られる.）変数名を適当につけてよいが，デフォールとでは variable になっている．

- エクセルの方が作業しやすいこともあるので，データセットを出力しておくとよい．
  - 標準メニュー＞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の順にクリック．
  - ポップアップが出てきたら，カンマ [,] のチェックを入れて OK ボタンを押す．
  - 出力される csv ファイル（Dataset.csv）を Excel で開く．変数名が左に 1 行ずれているので，修正しておく．このファイルの名前を practice3b として，xlsx 形式で保存しておくとよい．

エクセルにはランダムサンプリングの機能が備わっているが，説明は別の機会にする．そこで，ここでは各区分での平均値を trearment の欠損値に代入する．前章で記述した「並べ替え」や「フィルター」を使って，各区分のデータを取り出し，平均値を計算せよ．Excel では数式＞オート SUM の右側の印をクリックすると平均の関数が出てくるが，空いているセルに=average(V2:V100) とすると，V2 セルから V100 セルまでの平均が出力される．この平均が 0.5 よりも大きければ，その区分の欠損値を 1 とする．平均が 0.5 よりも小さければ，その区分の欠損値を 0 とする．このファイルの名前を practice3c とする．欠損値をすべて補正したファイルを読み込んで EZR で線形回帰を実行すると，10 分たっても計算が終わらない．諦めて，たとえば 2015 年から 2016 年の分だけで線形回帰することをお勧めする．（そのためのファイルが practice3d である．）

それを同じ区分の欠損値に代入し，データセット practice3c.xlsx を作成する．次に practice3c.xlsx を EZR で読み込み，線形回帰する．（Excel ではサンプルサイズが大き過ぎて回帰分析はできない．）統計解析＞連続変数の解析＞線形回帰（単回帰、重回帰）の順にクリックする．この後の手続きは前章で述べたとおりである．

ランダムサンプリングを実行する場合には，平均だけでなく，標準偏差も計算しておく．binomdist 関数で平均と標準偏差を指定すると，それに従って，0 か 1 かの値が選択される．