データ駆動型回帰分析 補足資料

前川 大空 *

2025 年 6 月 18 日

目次

1 回帰分析の課題

■データ駆動 従来の計量経済学では変数選択, ノンパラ, セミパラがこれにあたる. 何を指すのだろう · · ·

■一様妥当性 データ生成の母集団分布についての頑健性, といったことか. 頑健性と同じなのかしら · · ·

1.1 回帰分析

■p.1 観測可能性 観測可能性を前提としている. つまり, $\mathbf{W}$ はコントロール変数.

■p.1 構造モデル 構造モデルはデータ生成過程を表すのみ, 観測不可能な部分を誤差に全てまとめているため, 内生性などは排除されていない. つまり, 構造モデルは平均独立や条件付平均独立を満たすとは限らない.

■p.2 回帰モデル 回帰関数とは, 応答変数の条件付期待値関数のことを指す. 回帰モデルとは:

$$Y = g(S, \mathbf{W}) + e := \underbrace{\mathbb{E}[Y \mid S, \mathbf{W}]}_{\text{回帰関数}} + e$$

と回帰関数を (1.1) 式に適応させた式である. LIE から条件付平均独立 $\mathbb{E}[e \mid S, \mathbf{W}] = 0$ が確認できる.

■p.2 回帰関数の識別 回帰関数 $\mathbb{E}[Y \mid S, \mathbf{W}]$ は $(Y, S, \mathbf{W})$ の同時分布から一意に定まる. これは, 一般の分布について, 母集団モーメントは分布が判明することによって一意に定まるためである. 条件付分布 $Y|_{S,\mathbf{W}}$ の平均はこの特殊ケースと見なせる. ここで, $(Y, S, \mathbf{W})$ は全て観測可能である.

* 一橋大学経済学部 4 年, 五年一貫専修コース公共経済プログラム

Def: 識別

観測されるデータの同時分布が既知の時, $\boldsymbol{\theta}$ の値が一意に定まるならば, $\boldsymbol{\theta}$ は識別されるという.

上の定義から分かるように, 回帰関数は識別される. 何故ならば, 観測可能なデータ $(Y, S, \mathbf{W})$ の分布が既知の時, 上記の議論から回帰関数 $g(S, \mathbf{W}) := \mathbb{E}[Y \mid S, \mathbf{W}]$ は一意に定まるためである.

■p.2 **構造的/記述的な分析**　以下のような区別が為されている.

構造的/記述的な分析

構造的な分析: 何かしらの決定メカニズムを背後に想定する分析
記述的な分析: 観測される情報のみから識別可能な変数間の関係の分析

つまり, 回帰モデルによる分析は記述的な分析といえる.

■p.2 **構造モデルの識別**　回帰モデルでない構造モデル, つまり, $\mathbb{E}[e \mid S, \mathbf{W}] \neq 0$ である場合, g は回帰関数ではない別の関数になる. 関数が特定できないため, このままでは識別できない. 構造モデルへの追加的仮定は, 識別のために, 経済学ならば経済理論に基づいた妥当性が実証データからは検証できない仮定を置く. [*1]

■p.3 **構造モデルにおける誤差項の加法分離性**　(1.1) 式において, 大卒の因果効果を測るために $\mathbf{W}$ のみならず本来観測不可能な e も一定としていることに注意せよ. つまり, 因果効果の識別については, (1.1) 式に基づく構造的な分析の文脈においても述べられていない.

■p.3 **$\mathbf{W}$ 一定では高卒/大卒の賃金の差も一定という制約**　検証しておこう.

Proof 誤差項が加法分離可能な構造モデル:

$$Y = g(S, \mathbf{W}) + e \,,$$

について, $(\mathbf{w}_i, e_i)$ たる個人 i の教育年数の賃金への因果効果は,

$$Y_i \mid_{S_i=16} - Y_i \mid_{S_i=12} = [g(16, \mathbf{w}_i) + e_i] - [g(12, \mathbf{w}_i) + e_i] = g(16, \mathbf{w}_i) - g(12, \mathbf{w}_i) \,.$$

$\mathbf{w}_i = \mathbf{w}_j$ なる 2 個人の因果効果は $g(16, \mathbf{w}_i) - g(12, \mathbf{w}_i) = g(16, \mathbf{w}_j) - g(12, \mathbf{w}_j)$ で同一. □

■p.3 **回帰モデルと因果効果**　分かるのはあくまで平均で, 予測に過ぎない. 因果効果とは限らない.

■p.3 **因果推論は構造的な分析**　先述の例の通り, 因果効果は識別できるとは限らず, 単にメカニズムを記述したのみの, 即ち構造的な分析の範疇であった. しかし理論に基づく様々な仮定を置くことによって識別が可能となり, 因果効果の分析, 因果推論も記述的な分析に落とし込むことが出来る.

■p.4 **不均一分散の定義**　末石計量では, 無条件分散は説明変数が確率的な場合必ず定数になるため, 条件付分散を考えるのだ, との説明があったが, 本書では記述すら最早ない. 証明は末石計量の補足資料を参照のこと.

Def: 不均一分散

回帰モデルの仮定の下では無条件分散の不均一分散は実現せず, p.4 の形で定義を行う必要がある.

[*1] これが構造推定なる分野なのだろうか.

■p.4 **説明変数/応答変数**　語の用法として, 回帰モデルでない構造モデルには, この語を使うのは不適切?

■p.4 **限界効果　因果効果とは限らない**. 先述の通り, 回帰モデルによる分析は記述的な分析であって, 必ずしも構造的な分析だとは限らないためである. 1.1.2 章の内容は, 全体を通じて, 記述的な分析である, 回帰分析についての説明であることを理解しておかねばならない.

■p.4 **注 3 の内容**　『R による実証分析』は Rubin 流因果推論のフレームワークに基づく説明がなされていた. 具体的には, 記載されているように, Rubin の, 潜在結果モデルによって因果効果を定義していた. ここで重要なのは, (Rubin の) 因果効果は期待値の差分で定義されており, 一貫した関数の構造 $g(X_1, \mathbf{X}_{-1})$ を考える必要がないことだろう. この点で, Rubin の因果推論は, 本書での構造的な分析には当たらない. [*2] 一方で, 因果効果を不変の関数構造 $g(X_1, \mathbf{X}_{-1})$ の下での, 状態の変動による出力の変分として捉える点で, 本書は Pearl 流のアプローチ, ととらえればいいのだろうか.

■p.4 **因果効果としての限界効果**　回帰関数 μ が, 引数に対して一貫した構造を持つ際に, 限界効果は, はじめて因果効果としてみなすことが出来る.

■p.5 **コントロール変数**　p.1 での記述より, この回帰モデルの説明変数は全て観測可能である. 従って, 説明変数の一種であるコントロール変数にも観測可能性は必須であることには留意せよ.

■p.6 **回帰関数は MSPE の意味で最も良い応答変数の予測をもたらす**　コア計量等で証明したことがあるだろう. 『計量経済学のための数学』p.190 等にも証明が記載されている.

Proof 平均 2 乗予測誤差残差 (損失関数):

$$MSPE := \mathbb{E}[(Y - f(\mathbf{X}))^2] = \mathbb{E}[\varepsilon^2] \text{ where } \varepsilon = Y - f(\mathbf{X}),$$

の f による最小化問題を考えると, LIE と変形により:

$$\begin{aligned} MSPE &= \mathbb{E}[(Y - f(\mathbf{X}))^2] = \mathbb{E}[\mathbb{E}[\varepsilon^2 \mid \mathbf{X}]] \\ \varepsilon^2 &= \big[(Y - \mathbb{E}[Y \mid \mathbf{X}]) - (f(\mathbf{X}) - \mathbb{E}[Y \mid \mathbf{X}])\big]^2 \\ &= (Y - \mathbb{E}[Y \mid \mathbf{X}])^2 + (f(\mathbf{X}) - \mathbb{E}[Y \mid \mathbf{X}])^2 - 2(Y - \mathbb{E}[Y \mid \mathbf{X}])(f(\mathbf{X}) - \mathbb{E}[Y \mid \mathbf{X}]). \end{aligned}$$

ここで, ε^2 の各項について LIE のバリエーションを用いて変形し:

$$\begin{aligned} \mathbb{E}[\mathbb{E}[(\text{第一項}) \mid \mathbf{X}]] &= \mathbb{E}[\mathbb{E}[(Y - \mathbb{E}[Y \mid \mathbf{X}])^2 \mid \mathbf{X}]] = \mathbb{E}[\mathbb{E}[Y^2 \mid \mathbf{X}] - (\mathbb{E}[Y \mid \mathbf{X}])^2] \\ \mathbb{E}[\mathbb{E}[(\text{第二項}) \mid \mathbf{X}]] &= \mathbb{E}[\mathbb{E}[(f(\mathbf{X}) - \mathbb{E}[Y \mid \mathbf{X}])^2 \mid \mathbf{X}]] = \mathbb{E}[(f(\mathbf{X}) - \mathbb{E}[Y \mid \mathbf{X}])^2] \\ \mathbb{E}[\mathbb{E}[(\text{第三項}) \mid \mathbf{X}]] &= \mathbb{E}[\mathbb{E}[-2(Y - \mathbb{E}[Y \mid \mathbf{X}])(f(\mathbf{X}) - \mathbb{E}[Y \mid \mathbf{X}]) \mid \mathbf{X}]] \\ &= -2\mathbb{E}[(f(\mathbf{X}) - \mathbb{E}[Y \mid \mathbf{X}])\mathbb{E}[Y - \mathbb{E}[Y \mid \mathbf{X}] \mid \mathbf{X}]] \\ &= -2\mathbb{E}[(f(\mathbf{X}) - \mathbb{E}[Y \mid \mathbf{X}])\mathbb{E}[0]] = 0, \end{aligned}$$

以上を利用して, 以下の MSPE に関する不等式を得る:

$$\begin{aligned} MSPE = \mathbb{E}[\mathbb{E}[\varepsilon^2 \mid \mathbf{X}]] &= \mathbb{E}[\mathbb{E}[Y^2 \mid \mathbf{X}] - (\mathbb{E}[Y \mid \mathbf{X}])^2] + \mathbb{E}[(f(\mathbf{X}) - \mathbb{E}[Y \mid \mathbf{X}])^2] \\ &\geq \mathbb{E}[\mathbb{E}[Y^2 \mid \mathbf{X}] - (\mathbb{E}[Y \mid \mathbf{X}])^2] \ (f(\mathbf{X}) = \mathbb{E}[Y \mid \mathbf{X}] \text{ の時等号成立}). \end{aligned}$$

[*2] いわゆる, 『誘導的な分析』.

以上より, minimizer となる f が条件付期待値関数であることが分かった. □

■p.6 MSPE と MSE の別　MSE（Mean Squared Error）は, モデルの推定値 $\widehat{f}(X)$ が真の関数 $f(X)$ からどれだけずれているかを測る指標であり, 以下のように定義される:

$$MSE = \mathbb{E}\left[\left(\widehat{f}(X) - f(X)\right)^2\right]$$

一方, MSPE（Mean Squared Prediction Error）は, モデルの予測値 $\widehat{f}(X)$ と実際に観測された応答変数 Y とのずれを測るものであり, 次のように定義される:

$$MSPE = \mathbb{E}\left[\left(Y - \widehat{f}(X)\right)^2\right]$$

MSE は主にモデルの理論的な精度, 特にバイアスと分散に関する解析に用いられるのに対して, MSPE は未知データに対する予測性能の評価に使われる. ··· 少なくとも本書では. 『計量経済学のための数学』ではどうやら MPSE も MSE で書いている.

■p.6 バイアスと分散のトレードオフ　不偏性での用法と同じ『バイアス』だが, 不偏性があれば分散は大きくなるということ? 天気予報を常に晴れと予測するか, ちゃんと予測しようとするか, とかの例で言われる予測スコア? 情報量の話かなぁ. 未知のことの予測だから不偏推定量とかの話は出てこない?

■回帰分析の目的　まとめると以下の通り.

回帰分析の目的

1. 興味のある説明変数による, 応答変数への限界効果を調べること. 回帰モデルを構造モデルとして見なせるならば, これは因果効果の測定に他ならない.
2. 応答変数を予測すること.

1.2 線形回帰モデル

1.3 本書の課題と構成

1.4 補論

付録 A　記法

■不明点　分からない記述は赤文字を用いて記載する.

■ベクトル, 行列　共に `mathbf` を用いて記載する.

■条件付期待値　本文では $S = s$, $\mathbf{W} = \mathbf{w}$ である部分母集団について, $\mathbb{E}[Y \mid s, \mathbf{w}]$ と記述されている. 任意の S, $\mathbf{W}$ についてこの関係が成立する場合, $\mathbb{E}[Y \mid S, \mathbf{W}]$ と記載することにする.

■繰り返し期待値の法則　Law of Iterated Expectation, LIE と略す.

参考文献

[1] 末石 直也 (2024), データ駆動型回帰分析-計量経済学と機械学習の融合, 第 1 版, 日本評論社
[2] 末石 直也 (2015), 計量経済学 ミクロデータ分析へのいざない, 第 1 版, 日本評論社
[3] 星野 匡郎, 田中 久稔, 北川 梨津 (2023), R による実証分析: 回帰分析から因果分析へ, 第 2 版, オーム社
[4] 田中 久稔 (2019), 計量経済学のための数学, 第 1 版, 日本評論社