データ駆動型回帰分析 補足資料

前川 大空 *

2026 年 1 月 16 日

目次

* 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

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. データ先行の, 観測者の恣意性が入り込みうるモデルの特定化を避けることを本書の帯書きは強調している.

■一様妥当性 データ生成の母集団分布についての頑健性, といったところ.

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:

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

と回帰関数[*1] 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 $Y \mid_{S,\mathbf{W}}$ 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.

Def: 識別可能性

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 $g(S, \mathbf{W}) \coloneqq \mathbb{E}[Y \mid S, \mathbf{W}]$ 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. [*2]

[*1] 正確には, 回帰関数の内, OLS の下で Y の最良予測であるもの. 他の損失関数を選択すれば他の回帰関数が最良になりうる. 例えば, 最小絶対誤差法における最良の回帰関数は条件付中央値である. 詳しくは『計量経済学のための数学』の素敵な 9 章を参照.
[*2] 構造推定の役割の一つだろう.

■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, 因果効果の識別可能性については, **(1.1) 式** に基づく因果推論の文脈においても保証されていない.

■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,

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

$\mathbf{w}_i = \mathbf{w}_j$ 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. 特に e が加法分離可能なケースは直ちに記述的な分析に落とし込める. 必要なのは平均独立の仮定のみだ.

Proof **(1.1) 式** の e が加法分離可能な構造モデルを考える:

$$Y = h(D, \mathbf{X}) + e \,,$$

ここで, 観測可能な説明変数は $(D, \mathbf{X})$ であり, データが既知ならば LIE より以下のような形となる:

$$\mathbb{E}[Y \mid D, \mathbf{X}] = h(D, \mathbf{X}) + \mathbb{E}[e \mid D, \mathbf{X}] \,,$$

ここで, 回帰モデルの必要条件である平均独立の仮定 $\mathbb{E}[e \mid D, \mathbf{X}] = 0$ を置くと, $h(D, \mathbf{X})$ は回帰関数に等しくなり, 記述的な分析, さらに言えば回帰分析に落とし込むことが出来た. □

一般的な構造モデルにおいて要求される識別のための仮定は **1.4 章** にて言及される.

■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.

■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か.

■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, 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. [*3] 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 (暗黙にでも) 捉える点で, 経済学の文脈における因果推論で一般に採用されるのは, **構造モデルを念頭に置いた Rubin 流因果推論** である.

■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.5 **限界効果** 限界効果はコントロール変数の取り方によって変わる. 平均独立の仮定に違反しない限りは, コントロール変数は自由に選択可能で, この点が構造モデルを念頭に置いた因果効果との最大の違いだろう.

■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):

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

の f による最小化問題を考えると, LIE と変形により:

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

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:

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

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. □

[*3] いわゆる, 『誘導的な分析』.

■p.6 MSPE と MSE の別 MSE（Mean Squared Error）は, ある点 $(x_1, \ldots, x_p)$ におけるモデルの推定値 $\widehat{f}(\mathbf{x})$ が真の値 $f(\mathbf{x})$ 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:

$$\mathrm{MSE} = \mathbb{E}\left[\left(\widehat{f}\,(\mathbf{x}) - f\,(\mathbf{x})\right)^2\right]$$

一方, MSPE（Mean Squared Prediction Error）は, 母集団からのランダムサンプルについて, モデルの予測値 $\widehat{f}\,(\mathbf{X})$ 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:

$$\mathrm{MSPE} = \mathbb{E}\left[\left(Y - \widehat{f}\,(\mathbf{X})\right)^2\right] \tag{1.4}$$

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. [*4]

■p.6 バイアスと分散のトレードオフ 不偏性における『バイアス』とは, 任意の点においてではなく, ある点の近辺に限ったものである点で異なる. 以下の分解は, オーバーフィッティング (過学習) についてのトレードオフを説明するためのものである. 第二項は, $\widehat{f}$ が $\mathbf{x}$ に大きく左右されることを意味している.

$$\mathrm{MSE} = \underbrace{\left(\mathbb{E}[\widehat{f}\,(\mathbf{x})] - f\,(\mathbf{x})\right)^2}_{\text{バイアス}} + \underbrace{\mathrm{Var}\big[\widehat{f}\,(\mathbf{x})\big]}_{\text{分散}}$$

■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

■p.7 線形回帰モデル 線形回帰モデルとは, 回帰モデル:

$$Y = \mu\,(\mathbf{X}) + e \coloneqq \mathbb{E}[\,Y \mid \mathbf{X}\,] + e$$

において, 回帰関数 $\mathbb{E}[\,Y \mid \mathbf{X}\,]$ が実際に線形であることを仮定したモデルである. 即ち, 以下が成り立つ:

$$\mathbb{E}[\,Y \mid \mathbf{X}\,] = \mathbf{X}^\top \boldsymbol{\beta}, \quad \mathbb{E}[\,e \mid \mathbf{X}\,] \overset{\mathrm{LIE}}{=} 0$$

■p.8 $\boldsymbol{\beta}$ の推定 今, $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ と n 個の無作為標本が与えられている下で, $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_n]^\top$, $\mathbf{Y} = [Y_1, \ldots, Y_n]^\top$ との記法を利用して, 以下の関数の minimizer として OLS 推定値 $\widehat{\boldsymbol{\beta}}$ を決定する. [*5]

$$S(\mathbf{b}) \coloneqq \sum_{i=1}^{n} (Y_i - \mathbf{X}^\top \mathbf{b})^2 = (\mathbf{Y} - \mathbf{X}\mathbf{b})^\top (\mathbf{Y} - \mathbf{X}\mathbf{b}) = \|\mathbf{Y} - \mathbf{X}\mathbf{b}\|^2$$

$\mathbf{X}^\top \mathbf{X}$ が正則ならば, この最小化問題の解として OLS 推定量 $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$ は得られる.

[*4] ... 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.

[*5] 補足に記載したが, これは MS(P)E ではない. 他教科書でもこの目的関数自体には名前がついていないことが殆どだが, これは意図的なものだろう. あくまで推定量として適切な性質を持つものが計算の結果として出てくるために, この構成は正当化される.

Proof 散々コアコースで証明してきただろうが一応確認しておこう.

$$\begin{aligned} S(\mathbf{b}) &= (\mathbf{Y}-\mathbf{X}\mathbf{b})^\top(\mathbf{y}-\mathbf{X}\mathbf{b}) = (\mathbf{Y}^\top-\mathbf{b}^\top\mathbf{X}^\top)(\mathbf{Y}-\mathbf{X}\mathbf{b}) \\ &= \mathbf{Y}^\top\mathbf{Y}-\mathbf{b}^\top\mathbf{X}^\top\mathbf{Y}-\mathbf{Y}^\top\mathbf{X}\mathbf{b}+\mathbf{b}^\top\mathbf{X}^\top\mathbf{X}\mathbf{b} = \mathbf{Y}^\top\mathbf{Y}-2\mathbf{Y}^\top\mathbf{X}\mathbf{b}+\mathbf{b}^\top\mathbf{X}^\top\mathbf{X}\mathbf{b} \end{aligned}$$

最後の変形は第 $2,3$ 項がスカラーであるため. 正規方程式 (FONC) は以下のように変形できる:

$$\begin{aligned} \mathbf{0} = \frac{\partial S(\mathbf{b})}{\partial \mathbf{b}}\mid_{\mathbf{b}=\mathbf{b}_*} &= -2\frac{\partial}{\partial \mathbf{b}}\mathbf{Y}^\top\mathbf{X}\mathbf{b}\mid_{\mathbf{b}=\mathbf{b}_*}+\frac{\partial}{\partial \mathbf{b}}\mathbf{b}^\top\mathbf{X}^\top\mathbf{X}\mathbf{b}\mid_{\mathbf{b}=\mathbf{b}_*} \\ &= -2(\mathbf{Y}^\top\mathbf{X})^\top+(\mathbf{X}^\top\mathbf{X}+(\mathbf{X}^\top\mathbf{X})^\top)\mathbf{b}_* = -2\mathbf{X}^\top\mathbf{Y}+2\mathbf{X}^\top\mathbf{X}\mathbf{b}_* \end{aligned}$$

よって, $\mathbf{X}^\top\mathbf{Y}=\mathbf{X}^\top\mathbf{X}\mathbf{b}_*$, ここで $\mathbf{X}^\top\mathbf{X}$ が正則 (正定値) ならば, SOSC も満足し, さらに:

$$(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{X})\widehat{\boldsymbol{\beta}}=\widehat{\boldsymbol{\beta}}=(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}=\big(\sum_{i=1}^n\mathbf{X}_i\mathbf{X}_i^\top\big)^{-1}\sum_{i=1}^n\mathbf{X}_iY_i$$

と逆行列を利用できるため, 目的の表現を一意な最小解として得た. □

SOSC は **Thm 6.8** と $\mathbf{X}^\top\mathbf{X}$ の半正定値性から満たされる. $\mathbf{X}^\top\mathbf{X}$ の正則性は, 線形写像としての全単射を意味するため, 最小解の一意性を保証している.『計量経済学のための数学』では興味の対象を推定するための適切な性質 (一致性) があることを確認したのちにこれを OLSE と定義していたが, 本書をはじめ多くの本は, この定義から出発して推定量としての正統性を探る. 詳しくは **付録 B.1** を確認すること.

■p.8 $\mathbf{X}^\top\mathbf{X}$ **の正則性** 言い換えれば, $\mathbf{X}$ は列フルランク, 説明変数間に完全な多重共線性がない状態である. 上記の証明を見ればわかる通り, これは OLSE の一意性を保証するために必要となる条件である. この成立のためには $n\geq k$ が必要条件である. 詳しい議論は **付録 B.2** を見よ. 『古典的な統計学の設定』においては $n\gg k$ が想定され, この場合列フルランク性が崩れることはめったにないため, 重回帰モデルの仮定にこれを課すことは殆ど問題を引き起こさないわけだが, k が大きい, さらには $n<k$ となってしまってはこの仮定はもはや満たされなくなってしまう. この問題は, 正則化などを扱う教科書後半で議論されることだろう.

■p.9 OLSE **の不偏性** 末石計量とは少し違う流れ. 今回は『データ駆動型』に沿って確認してみよう.

Proof まず OLSE と回帰係数 $\boldsymbol{\beta}$ の関係性を確認しておく:

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}+\mathbf{e}) \\ &= \boldsymbol{\beta}+(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{e} = \boldsymbol{\beta}+\big(\sum_{i=1}^n\mathbf{X}_i\mathbf{X}_i^\top\big)^{-1}\sum_{i=1}^n\mathbf{X}_ie_i \end{aligned}$$

さて, 変形した式を用いて不偏性を確認しよう. まず, LIE により $\mathbb{E}[\widehat{\boldsymbol{\beta}}]=\mathbb{E}[\mathbb{E}[\widehat{\boldsymbol{\beta}}\mid\mathbf{X}]]$ が成り立ち:

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}\mid\mathbf{X}]=\boldsymbol{\beta}+(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbb{E}[\mathbf{e}\mid\mathbf{X}]=\boldsymbol{\beta} \tag{1.7}$$

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}]=\mathbb{E}[\mathbb{E}[\widehat{\boldsymbol{\beta}}\mid\mathbf{X}]]=\mathbb{E}[\boldsymbol{\beta}]=\boldsymbol{\beta}$$

以上から目的の関係を得る. なお, 上記の計算に利用した $\mathbb{E}[\mathbf{e} \mid \mathbf{X}] = \mathbf{0}$ は

$$\mathbb{E}[\mathbf{e} \mid \mathbf{X}] = \begin{pmatrix} \mathbb{E}[e_1 \mid \mathbf{X}] \\ \vdots \\ \mathbb{E}[e_n \mid \mathbf{X}] \end{pmatrix} = \begin{pmatrix} \mathbb{E}[\mathbb{E}[e_1 \mid \mathbf{X}_1] \mid \mathbf{X}] \\ \vdots \\ \mathbb{E}[\mathbb{E}[e_n \mid \mathbf{X}_n] \mid \mathbf{X}] \end{pmatrix} = \begin{pmatrix} \mathbb{E}[0 \mid \mathbf{X}] \\ \vdots \\ \mathbb{E}[0 \mid \mathbf{X}] \end{pmatrix} = \mathbf{0}$$

と LIE を利用することにより正当性が確かめられる. □

■p.9 OLSE **の条件付共分散行列 (1.8) 式** まで導出しよう.

Proof まず, $\mathbb{E}[\widehat{\boldsymbol{\beta}} \mid \mathbf{X}] = \boldsymbol{\beta}$ を満たす推定量の条件付共分散行列は以下のように簡単に書ける:

$$\mathrm{Var}\,(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}) \coloneqq \mathbb{E}\big[\big(\widehat{\boldsymbol{\beta}} - \mathbb{E}[\widehat{\boldsymbol{\beta}} \mid \mathbf{X}]\big)\big(\widehat{\boldsymbol{\beta}} - \mathbb{E}[\widehat{\boldsymbol{\beta}} \mid \mathbf{X}]\big)^\top \mid \mathbf{X}\big] = \mathbb{E}\big[\big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\big)\big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\big)^\top \mid \mathbf{X}\big]$$

これと不偏性での計算を利用して以下を得る:

$$\begin{aligned} \mathrm{Var}\,(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}) &= \mathbb{E}\big[\big((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{e}\big)\big((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{e}\big)^\top \mid \mathbf{X}\big] \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbb{E}\big[\mathbf{e}\mathbf{e}^\top \mid \mathbf{X}\big]\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\ &= \big(\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top\big)^{-1}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top\mathbb{E}\big[e_i^2 \mid \mathbf{X}_i\big]\big(\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top\big)^{-1} \end{aligned}$$

最後の変形は **回帰モデルでの誤差項の独立性** から, $\mathbb{E}\big[\mathbf{e}\mathbf{e}^\top \mid \mathbf{X}\big]$ が対角行列になることを利用した:

$$\mathbb{E}[\mathbf{e}\mathbf{e}^\top \mid \mathbf{X}] = \begin{bmatrix} \mathbb{E}[e_1^2 \mid \mathbf{X}] & \cdots & \mathbb{E}[e_1e_n \mid \mathbf{X}] \\ & \ddots & \\ \mathbb{E}[e_ne_1 \mid \mathbf{X}] & \cdots & \mathbb{E}[e_n^2 \mid \mathbf{X}] \end{bmatrix} \overset{\text{i.i.d.}}{=} \begin{bmatrix} \mathbb{E}[e_1^2 \mid \mathbf{X}] & \cdots & \mathbf{O} \\ & \ddots & \\ \mathbf{O} & \cdots & \mathbb{E}[e_n^2 \mid \mathbf{X}] \end{bmatrix}$$

均一分散 $\mathrm{Cov}(\mathbf{e} \mid \mathbf{X}) = \sigma^2\mathbf{I}_n$ を仮定すれば:

$$\mathrm{Var}\,(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}) = \sigma^2\,(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} = \sigma^2\,(\mathbf{X}^\top\mathbf{X})^{-1} = \sigma^2\,\big(\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top\big)^{-1} \tag{1.8}$$

よって目標の式を得られた. □

■p.11 t **統計量の一様妥当性** 標本平均について t (検定) 統計量を構成する中でこれを確かめてみよう. 手始めに $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_X, \sigma_X^2)$ を考えてみよう. 先ず, σ_X^2 が既知として, $H_0 = \mu_{X,0}$, 有意水準 $1-\alpha$ の両側検定を考える. 実際に帰無仮説が正しければ, t 統計量は:

$$T = \sqrt{n}\frac{\bar{X_n} - \mu_{X,0}}{\sigma_X} \sim \mathcal{N}(0,1)$$

で標準正規分布に従う. 次に, σ_X^2 も未知として同様の両側検定を考える. ここで, t 統計量は:

$$T = \sqrt{n}\frac{\bar{X_n} - \mu_{X,0}}{s_X} = \frac{\sqrt{n}\frac{\bar{X_n} - \mu_{X,0}}{\sigma_X}}{\sqrt{\frac{(n-1)s_X^2/\sigma_X^2}{n-1}}} \sim t(n-1)$$

で, 自由度 $n-1$ の t 分布に従う. さらに一般化し, $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim}$, $\mathbb{E}[X_i^4] < \infty$ とする. ここで t 統計量は, (分母) は 1 に確率収束し, (分子) は $\mathcal{N}(0,1)$ に分布収束する ($\because$ CLT). Slutsky's Theorem より:

$$T = \sqrt{n}\frac{\bar{X}_n - \mu_{X,0}}{s_X} \xrightarrow{d} \mathcal{N}(0,1)$$

標本サイズが大きい限り, X の分布に関わらず標準正規分布で良く近似できることが分かった. OLSE $\widehat{\boldsymbol{\beta}}$ の漸近正規性から, 最後の部分と同様の議論が OLSE に関しても繰り返すことが出来る.

■p.11 **線形射影モデル** 線形射影モデルの推定対象は線形関数の中での MSPE の minimizer だが, 回帰モデルは回帰関数の関数形を線形に特定してから, MSPE の minimizer を推定していることに注意せよ.

線形射影モデル: $Y = \mathbf{X}^\top \boldsymbol{\beta} + e \quad \text{where } \boldsymbol{\beta} = \arg\min_{\mathbf{b}} \mathbb{E}[(Y - \mathbf{X}^\top \mathbf{b})^2]$

線形回帰モデル: $Y = \mu(\mathbf{X}) + e \coloneqq \mathbb{E}[Y \mid \mathbf{X}] + e \quad \text{where } \mathbb{E}[Y \mid \mathbf{X}] = \mathbf{X}^\top \boldsymbol{\beta}$

■p.11 **線形射影係数** 線形射影モデルの識別対象は線形射影係数だが, これを表す **(1.11) 式** を導出する:

Proof まず, 目的関数を二次形式として明示的に書こう:

$$\mathbb{E}\big[(Y - \mathbf{X}^\top \mathbf{b})^2\big] = \mathbb{E}[Y^2] - 2\,\mathbf{b}^\top \mathbb{E}[\mathbf{X}Y] + \mathbf{b}^\top \mathbb{E}[\mathbf{X}\mathbf{X}^\top]\,\mathbf{b}.$$

以下の FONC は $\mathbb{E}[\mathbf{X}\mathbf{X}^\top]$ が仮定されたその正則性より正定値であることから SOSC も満足し:

$$-2\,\mathbb{E}[\mathbf{X}Y] + 2\,\mathbb{E}[\mathbf{X}\mathbf{X}^\top]\,\mathbf{b} = \mathbf{0}, \quad \mathbb{E}[\mathbf{X}\mathbf{X}^\top]\,\boldsymbol{\beta} = \mathbb{E}[\mathbf{X}Y]$$

以上から, 線形射影モデルで識別の対象となる線形射影係数は以下のように表現される:

$$\boldsymbol{\beta} = \big(\mathbb{E}[\mathbf{X}\mathbf{X}^\top]\big)^{-1}\,\mathbb{E}[\mathbf{X}Y] \tag{1.11}$$

また, 定義より $e = Y - \mathbf{X}^\top \boldsymbol{\beta}$ で, FONC から:

$$\mathbb{E}[\mathbf{X}e] = \mathbb{E}\big[\mathbf{X}\,(Y - \mathbf{X}^\top \boldsymbol{\beta})\big] = \mathbb{E}\big[\mathbf{X}Y\big] - \mathbb{E}\big[\mathbf{X}\mathbf{X}^\top\big]\boldsymbol{\beta} = \mathbf{0} \tag{1.10}$$

この式は残差 $\widehat{\mathbf{e}}$ が $\mathbf{X}$ の直交補空間に属することに対応した結果である. □

詳しくは **付録 B.1** を確認すること.

■p.11 **線形射影と回帰関数** $\mathbf{X}^\top \boldsymbol{\beta}$ は回帰関数 $\mu(\mathbf{X}) = \mathbb{E}[Y \mid \mathbf{X}]$ の最良線形近似でもある.

Proof 線形射影係数の定義は以下のように変形できる:

$$\begin{aligned}
\boldsymbol{\beta} &= \arg\min_{\mathbf{b}} \mathbb{E}[(Y - \mathbf{X}^\top \mathbf{b})^2] = \arg\min_{\mathbf{b}} \mathbb{E}[(Y - \mu(\mathbf{X}) + \mu(\mathbf{X}) - \mathbf{X}^\top \mathbf{b})^2] \\
&= \arg\min_{\mathbf{b}} \mathbb{E}[(Y - \mu(\mathbf{X}))^2 + (\mu(\mathbf{X}) - \mathbf{X}^\top \mathbf{b})^2 + 2(Y - \mu(\mathbf{X}))(\mu(\mathbf{X}) - \mathbf{X}^\top \mathbf{b})] \\
&= \arg\min_{\mathbf{b}} \mathbb{E}[(Y - \mu(\mathbf{X}))^2] + \mathbb{E}[(\mu(\mathbf{X}) - \mathbf{X}^\top \mathbf{b})^2] + 2\mathbb{E}[(Y - \mu(\mathbf{X}))(\mu(\mathbf{X}) - \mathbf{X}^\top \mathbf{b})] \\
&\overset{\text{LIE}}{=} \arg\min_{\mathbf{b}} \mathbb{E}[(Y - \mu(\mathbf{X}))^2] + \mathbb{E}[(\mu(\mathbf{X}) - \mathbf{X}^\top \mathbf{b})^2] + 2\mathbb{E}[(\mu(\mathbf{X}) - \mu(\mathbf{X}))(\mu(\mathbf{X}) - \mathbf{X}^\top \mathbf{b})] \\
&= \arg\min_{\mathbf{b}} \mathbb{E}[(Y - \mu(\mathbf{X}))^2] + \mathbb{E}[(\mu(\mathbf{X}) - \mathbf{X}^\top \mathbf{b})^2] = \arg\min_{\mathbf{b}} \mathbb{E}[(\mathbb{E}[Y \mid \mathbf{X}] - \mathbf{X}^\top \mathbf{b})^2]
\end{aligned}$$

(1.11) 式 も以下のように LIE を利用して変形できる:

$$\boldsymbol{\beta} = \left(\mathbb{E}[\mathbf{X}\mathbf{X}^\top]\right)^{-1} \mathbb{E}[\mathbf{X}\mathbb{E}[Y \mid \mathbf{X}]]$$

目的関数と結果がそっくり Y から $\mu(\mathbf{X})$ に入れ替わった形ゆえ, 主張の正統性が確認できた. □

モデルの区別

線形回帰モデル: 平均的な関係を示す. $\mathbb{E}[Y \mid \mathbf{X}] = \mathbf{X}^\top\boldsymbol{\beta}$ と回帰関数が線形で, $\mathbb{E}[e \mid \mathbf{X}] = 0$
線形射影モデル: 回帰関数の関数形に関わらない線形最良近似, $\mathbb{E}[\mathbf{X}e] = \mathbf{0}, \mathbb{E}[e] = 0$
内生性のある構造型: $\mathrm{Cov}(\mathbf{X}_i, e_i) \neq \mathbf{0}$ で上記どちらの仮定にも違反しうるが, 生成過程を記述する

■p.13 **欠落変数バイアス**　以下のような図示が可能だ. 特に特定の限界効果に関心があれば説明変数を過不足なくモデルに入れ込む必要があるが, 能力等の観測不可能な説明変数はこれを不可能にしてしまう. 先述のとおり, 限界効果は回帰関数の差分または微分係数に過ぎず, 説明変数の取り方によっていくらでも変わりうる.

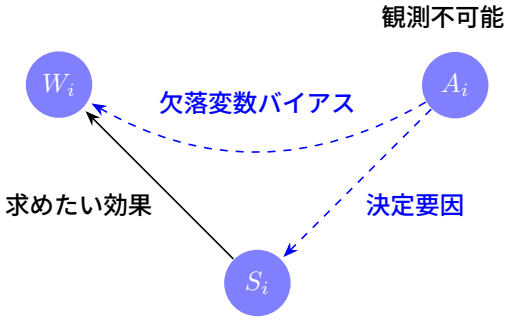

図 1　欠落変数バイアス

1.3　本書の課題と構成

■p.15 **回帰関数の線形性の仮定が不適切な可能性**　言い換えれば, 推定している線形射影モデルがもはや線形回帰モデルと一致しない状況で引きおこる問題点である, 先述の通り, 線形射影は回帰関数に対する最良線形近似を与えはするが, それは $\mathbf{X}$ の全域に対しての平均的な意味であり, ある実現値 $\mathbf{X} = \mathbf{x}_0$ において適切な予測値を返してくれるとは限らない. 以下の図示によって理解できる. MSE と MSPE の差異も確認できる.

■p.15 **回帰変数が多いほど推定にまつわる分散が大きくなる**　過学習の話だろうか. **4 章** で扱われる『次元の呪い』と同じ意味の語だろうか.

■p.15 **高次元データ**　以下のように記載がある.

Def: 高次元データ

サンプルサイズより共変量の数が大きいデータ

先述の通り, もはや高次元データでは $\mathbf{X}^\top\mathbf{X}$ の正則性が満たされず, OLSE が一意に定まらない. **6 章** で扱われる正則化がこの解決策のひとつとして予見される.

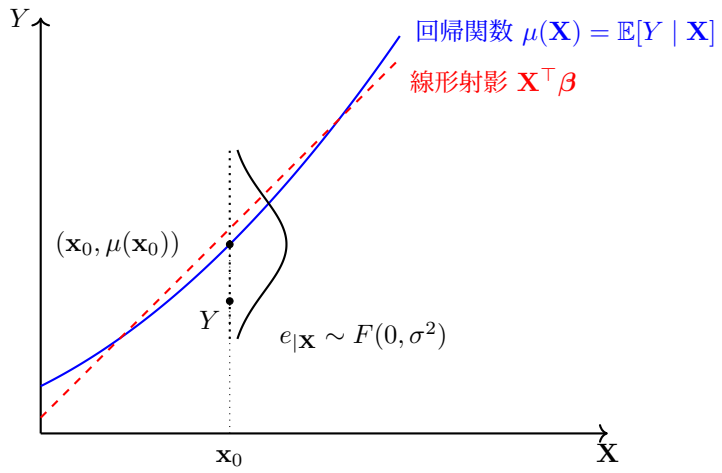

図 2　非線形な回帰関数 $\mu(\mathbf{X})$ とその線形射影

■p.17 **変数選択　2 章** では変数選択を行った後に統計的推測を行うことの問題点を論じるようだが, このいくつかの高次元データにおける解決策は **7 章** で示される.

1.4　補論

■p.18 SUTVA　数学的にはたったこれだけの話, 常時暗黙に仮定されていることに留意せよ:

Assumption: SUTVA

潜在的結果が以下のように表されることを指す.

$$Y_i = Y_i(D_i) = \begin{cases} Y_i(1) & \text{if } D_i = 1 \\ Y_i(0) & \text{if } D_i = 0 \end{cases} \quad \text{where } D_i \in \{0, 1\}$$

つまり, 以下の 2 条件がこの式で表現されている:

1. 潜在的結果は他者の処置の影響を受けない (引数が D_i のみ)
2. 処置が均一 (二値変数, デジタルな処置)

■p.18 **構造モデルの導入**　構造モデルの建付けからも分かるように, 関数 h の形状自体は人によって, そして処置によって変わらないことに注意せよ. SUTVA を満たしていることは容易に確認できるだろう. **p.3** の議論を踏まえてか, e は加法分離可能ではない.

■p.19 **構造モデルの処置効果　(1.14) 式** の構造モデルにおける Rubin の意味での『処置効果』は:

$$Y_i = h\left(D_i, \mathbf{X}_i, e_i\right) \tag{1.14}$$

$$Y_i\left(1\right) - Y_i\left(0\right) = h\left(1, \mathbf{X}_i, e_i\right) - h\left(0, \mathbf{X}_i, e_i\right) \tag{1.15}$$

しかしこの場合は限界効果とは言えない. これを, 記述的な分析でないことの確認をもって確かめよう.

Proof **(1.14) 式** の形の構造モデルを考える:

$$Y_i = h\left(D_i, \mathbf{X}_i, e_i\right)$$

ここで, 観測可能な説明変数は $(D_i, \mathbf{X}_i)$ であり, データが既知ならば以下のような形となる:

$$\mathbb{E}[\, Y_i \mid D_i, \mathbf{X}_i \,] = \mathbb{E}[\, h\left(D_i, \mathbf{X}_i, e_i\right) \mid D_i, \mathbf{X}_i \,]$$

e_i は関数 h に加法分離不可能な形で入るため, 最早 (単純には) LIE は適応不能. 右辺は観測不能な変数が入っているため, 識別不可能であり, 従って追加的な仮定を置かない限り記述的な分析でない. □

■p.19 Rubin **の因果推論の利点** $Y_i(D_i)$ のまま, データ生成過程を特定せずに因果推論が可能となること.

■p.19 **経済学における『処置効果』** 多くの場合, 経済学における『処置効果』は, Rubin の意味でも, 本書の構造的な分析における因果効果の意味でも, さらに回帰モデルである構造モデルの (二値変数の場合の) 限界効果の意味においても当てはまる. 特に e が加法分離可能な構造モデルはこれに当てはめられる.

Proof 再び, (1.1) 式の e が加法分離可能な構造モデルを考える:

$$Y_i = h\left(D_i, \mathbf{X}_i\right) + e_i$$

D_i はダミー変数 (二値変数) で, 暗黙に SUTVA は満たされる. ここで, 本書の因果効果は:

$$Y_i \mid_{D_i=1} - Y_i \mid_{D_i=0} = h\left(1, \mathbf{X}_i\right) - h\left(0, \mathbf{X}_i\right) \tag{$*$}$$

一方で Rubin の因果効果も以下のように表現できる:

$$Y_i\left(1\right) - Y_i\left(0\right) = h\left(1, \mathbf{X}_i\right) - h\left(0, \mathbf{X}_i\right) \tag{$**$}$$

先述の通り, 回帰モデルの必要条件である平均独立の仮定 $\mathbb{E}[\, e_i \mid D_i, \mathbf{X}_i \,] = 0$ を置くと, $h\left(D_i, \mathbf{X}_i\right)$ は回帰関数に等しくなり, 記述的な分析, さらに言えば回帰分析に落とし込むことが出来た. 限界効果は:

$$\mathbb{E}[\, Y_i \mid 1, \mathbf{X}_i \,] - \mathbb{E}[\, Y_i \mid 0, \mathbf{X}_i \,] = h\left(1, \mathbf{X}_i\right) - h\left(0, \mathbf{X}_i\right) \tag{$***$}$$

ここで $(*)$, $(**)$, $(***)$ は全て同一であることから, 上の記載に合致した. □

■p.19 Rubin **流因果推論の識別対象** まず, 分析間の差異を整理する.

構造分析: $Y_i = h\left(D_i, \mathbf{X}_i, e_i\right),$ where $h\left(D_i = 1, \mathbf{X}_i, e_i\right) - h\left(D_i = 0, \mathbf{X}_i, e_i\right)$: interested
回帰分析: $Y_i = \mathbb{E}[Y_i \mid D_i, \mathbf{X}_i] + e_i,$ where $\mathbb{E}[Y_i \mid D_i = 1, \mathbf{X}_i] - \mathbb{E}[Y_i \mid D_i = 0, \mathbf{X}_i]$: interested
反実仮想フレームワーク: $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0),$ where $Y_i(1) - Y_i(0)$: interested

この三つ (仮定によっては同一視できうる) について, 何が観測不可能なのかは検討する必要がある.

観測不可能性

フレームワークによって何が観測不可能なのか, それに伴い生起する性質は異なる:

構造分析： 誤差項に入れ込まれる変数, 内生性の原因
回帰分析： 回帰モデルに入れたい変数の内いくつか, 欠落変数バイアスの原因
反実仮想： $(Y_i(0), Y_i(1))$ の一方（『因果推論の根本問題』）

Def: ATE, ATET

『因果推論の根本問題』から, 個人の処置効果ではなく, 以下の対象の識別をすることが要求される.

$$\tau_{\text{ATE}} \coloneqq \mathbb{E}[Y_i(1) - Y_i(0)] \tag{1.16}$$
$$\tau_{\text{ATET}} \coloneqq \mathbb{E}[Y_i(1) - Y_i(0) \mid D_i = 1] \tag{1.17}$$

■p.19 **セレクションバイアス** 主体は意思決定の中で処置を選び取る. 意思決定モデルを考えたとき, 処置はこの中で選び取られる変数, との意味で内生変数である. 内生変数は一般に, 応答変数との間に, 交絡変数を通じて相関を持つ. 構造モデルとして因果推論を考えた時, 観測不可能な交絡変数 (共変量) は誤差項としてまとめられる. 結果的に, 処置は誤差項との間に相関を持つが, この状態こそ内生性の定義である. さらに, 処置が誤差項との相関を通じて潜在的結果と相関を持っており, これがセレクションバイアスと呼ばれるものである.

■p.20 **交絡変数** Rubin 流の因果推論の文脈で定義される.

Def: 交絡変数

潜在的結果と処置の両方に影響を与えるような変数

■p.21 **強い無視可能性** 識別は観測されるデータについての概念で, 今回は以下のように分布を分類できる:

観測可能: $f_{Y(1)|D}(y \mid 1),\ f_{Y(0)|D}(y \mid 0)$
観測不可能: $f_{Y(1)|D}(y \mid 0),\ f_{Y(0)|D}(y \mid 1),\ f_{Y(1)}(y),\ f_{Y(0)}(y)$

識別のため, つまり推定対象 $(\tau_{\text{ATE/T}})$ を全て観測可能なもので表すための仮定として, 以下が課される:

Assumption: 強い無視可能性

1. 非交絡の仮定, 条件付独立の仮定:

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp D_i \mid \mathbf{X}_i$$

2. オーバーラップの仮定:

$$0 < \mathbf{P}\{D_i = 1 \mid \mathbf{X}_i = \mathbf{x}\} < 1 \ \forall \mathbf{x}$$

■p.20 **非交絡の仮定の意味** 非交絡の仮定は, 構造モデル **(1.14) 式** を想定すると以下に対応する:

$$f(e \mid D, \mathbf{X}) = f(e \mid \mathbf{X})$$

ここで f は e の条件付分布. これを示したい. 三輪くんが示してくれた方法を真似してみよう.

Proof 確率空間 $(\Omega, \mathcal{F}, \mathbf{P})$ 上で, 確率及び分布は全て $\mathbf{X}$ の条件付とする. **(1.14) 式** の構造モデル:

$$Y \coloneqq h\big(D,\ \mathbf{X},\ e\big)$$

を考える. 確率変数ベクトル $\mathbf{Z} \coloneqq (D, e)^\top : \Omega \to \{0,1\} \times \mathbb{R}$ を定義すると, 確率変数 $Y = h\big(D,\ \mathbf{X},\ e\big)$ は所与の $\mathbf{X}$ の下で完全に $\mathbf{Z}$ で決定されるため, $\mathbf{Z}$ **可測** である. すなわち以下の包含関係が成立する:

$$\sigma\big[Y\big] \subset \sigma\big[\mathbf{Z}\big] \subset \mathcal{F}$$

ここで $\sigma[Y]$ は **確率変数 Y が生成する σ-加法族** を意味する:

$$\sigma\big[Y\big] \coloneqq \sigma\big[\{\{Y \leq t\} \mid t \in \mathbb{R}\}\big] \subset \mathcal{F}$$

任意の $(x, y) \in \mathbb{R}^2$ で, e が適切な条件を満たす値域の範囲を以下のように表現できる:

$$A_{x,y} = \{e \mid Y(1) \leq x, Y(0) \leq y\} = \{e \mid h(1, \mathbf{X}, e) \leq x, h(0, \mathbf{X}, e) \leq y\} \in \mathcal{B}$$

ここで, $\mathcal{B}$ は e の値域 $\mathbb{R}$ 上の **ボレル集合族** である. よって非交絡の仮定:

$$(Y\,(1),\, Y\,(0)) \perp\!\!\!\perp D \mid \mathbf{X}$$

は確率変数の $\mathbf{X}$ の条件付独立を意味する. まず事象 $A_1, \ldots, A_n \in \mathcal{F}$ の独立性とは以下を指す:

$$\mathbf{P}\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} \mathbf{P} A_i$$

部分事象族 $\mathcal{G}_1, \ldots, \mathcal{G}_n \subset \mathcal{F}$ が独立とは, 任意の事象 $A_1 \in \mathcal{G}_1, \ldots, A_n \in \mathcal{G}_n$ がつねに独立になることをいう. 確率変数 $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ が独立であるとは, $\sigma[\mathbf{Z}_1], \ldots, \sigma[\mathbf{Z}_n]$ が独立であることを指す. これらの定義より, **略記** を利用することで非交絡の仮定は以下のように表現できる:

$$\begin{aligned}
\mathbf{P}\big(\{D=0\} \cap \{e \in A_{x,y}\}\big) &= \mathbf{P}\{e \in A_{x,y}\}\, \mathbf{P}\{D=0\} \\
&= \mathbf{P}\big(\{D \in \{0,1\}\} \cap \{e \in A_{x,y}\}\big)\, \mathbf{P}\big(\{D=0\} \cap \{e \in \mathbb{R}\}\big)\,, \\
\mathbf{P}\big(\{D=1\} \cap \{e \in A_{x,y}\}\big) &= \mathbf{P}\{e \in A_{x,y}\}\, \mathbf{P}\{D=1\} \\
&= \mathbf{P}\big(\{D \in \{0,1\}\} \cap \{e \in A_{x,y}\}\big)\, \mathbf{P}\big(\{D=1\} \cap \{e \in \mathbb{R}\}\big)\,.
\end{aligned}$$

ここで $h(D, \mathbf{X}, e)$ が e について単調増加と仮定すると, $h(d, \mathbf{X}, e),\ d \in \{0,1\}$ には e についての逆関数がそれぞれ存在し, これを $h_d^{-1}(\cdot)$ と書くと, $A_{x,y}$ は以下のように書き換えられる:

$$A_{x,y} = \{\, e \mid e \leq h_1^{-1}(x),\ e \leq h_0^{-1}(y)\} = (-\infty, \min\{h_1^{-1}(x), h_0^{-1}(y)\}]\,,$$

従って $d = 0, 1$ について:

$$\begin{aligned}
&\mathbf{P}\big(\{D=d\} \cap \{e \in (-\infty, \min\{h_1^{-1}(x), h_0^{-1}(y)\}]\}\big) \\
&\qquad = \mathbf{P}\{e \in (-\infty, \min\{h_1^{-1}(x), h_0^{-1}(y)\}]\}\, \mathbf{P}\{D=d\}
\end{aligned}$$

オーバーラップの仮定のもとでは $\mathbf{P}\{D=d\} > 0$ であるから, 両辺をその値で割って:

$$\frac{\mathbf{P}\big(\{D=d\} \cap \{e \in (-\infty, \min\{h_1^{-1}(x), h_0^{-1}(y)\}]\}\big)}{\mathbf{P}\{D=d\}} = \mathbf{P}\{e \in (-\infty, \min\{h_1^{-1}(x), h_0^{-1}(y)\}]\} \qquad (*)$$

x, y を任意にとれば $\min\{h_1^{-1}(x), h_0^{-1}(y)\}$ も任意にとれるので，f を e の条件付分布として

$$f(e \mid D, \mathbf{X}) = f(e \mid \mathbf{X})$$

が成り立つ．オーバーラップの仮定が成り立たない場合は $\mathbf{P}\{D = d\} = 0$ なる d について条件付き確率は定義されないので考慮不要．また $h(D, \mathbf{X}, e)$ が e について単調減少のときは，$(*)$ 式にて $A_{x,y}$ を $[\max\{h_1^{-1}(x), h_0^{-1}(y)\}, \infty)$ に置き換え，両辺を 1 から引けば同様に示される． □

$h(D, \mathbf{X}, e)$ の e についての単調性が証明には必要なことが分かった．

■p.21 **強い無視可能性による識別** ATE, ATET は強い無視可能性によって識別が可能になる．

Theorem: ATE, ATET の識別

強い無視可能性のもとで, ATE, ATET は以下のように識別可能な形で表現される:

$$\begin{aligned}\tau_{\mathrm{ATE}} &\coloneqq \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0], \\ \tau_{\mathrm{ATET}} &\coloneqq \mathbb{E}[\, Y_i(1) - Y_i(0) \mid D_i = 1\,] = \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[\mathbb{E}[Y_i \mid D_i = 0, \mathbf{X}_i] \mid D_i = 1]\end{aligned}$$

Proof まず ATE について考える．以下のように $\tau(\mathbf{x})$ を定義する:

$$\tau(\mathbf{x}) \coloneqq \mathbb{E}[Y_i(1) - Y_i(0) \mid \mathbf{X}_i = \mathbf{x}]$$

強い無視可能性により, 任意の $\mathbf{x}$ で:

$$\mathbb{E}[Y_i(d) \mid \mathbf{X}_i = \mathbf{x}] \overset{\text{非交絡}}{=} \mathbb{E}[Y_i(d) \mid D_i = d, \mathbf{X}_i = \mathbf{x}] \overset{\text{def}}{=} \mathbb{E}[Y_i \mid D_i = d, \mathbf{X}_i = \mathbf{x}] \quad (d \in \{0, 1\})$$

が成り立ち, これは観測可能な量である．よって $\tau(\mathbf{x})$ も観測可能な量で以下のように表記できる:

$$\begin{aligned}\tau(\mathbf{x}) &= \mathbb{E}[Y_i(1) \mid \mathbf{X}_i = \mathbf{x}] - \mathbb{E}[Y_i(0) \mid \mathbf{X}_i = \mathbf{x}] \\ &= \mathbb{E}[Y_i \mid D_i = 1, \mathbf{X}_i = \mathbf{x}] - \mathbb{E}[Y_i \mid D_i = 0, \mathbf{X}_i = \mathbf{x}]\,,\end{aligned}$$

(Y_i, D_i) は観測可能で, この同時分布が判明すれば条件付期待値は一意に定まるため, 任意の $\mathbf{x}$ に対して $\tau(\mathbf{x})$ は実際に識別できる．ここで, LIE から $\tau(\mathbf{x})$ は識別の対象 τ_{ATE} と以下の関係を持つ:

$$\tau_{\mathrm{ATE}} \coloneqq \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[\mathbb{E}[Y_i(1) - Y_i(0) \mid \mathbf{X}_i]] = \mathbb{E}[\tau(\mathbf{X})]$$

期待値は $\mathbf{X}$ について取られる．観測可能なもので全て記載できているため, τ_{ATE} も識別可能である．

$$\tau_{\mathrm{ATE}} = \mathbb{E}[\tau(\mathbf{X}_i)] \overset{\text{LIE}}{=} \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]\,.$$

τ_{ATET} も $\tau(\mathbf{x})$ を利用して以下のように表現できる:

$$\tau_{\mathrm{ATET}} \coloneqq \mathbb{E}[Y_i(1) - Y_i(0) \mid D_i = 1] = \mathbb{E}[\mathbb{E}[Y_i(1) - Y_i(0) \mid D_i = 1, \mathbf{X}_i] \mid D_i = 1] = \mathbb{E}[\tau(\mathbf{X}_i) \mid D_i = 1]$$

ここでも, 期待値は $\mathbf{X}$ について取られている．強い無視可能性により, 任意の $\mathbf{x}$ について:

$$\mathbb{E}[Y_i(d) \mid \mathbf{X}_i = \mathbf{x}] = \mathbb{E}[Y_i(d) \mid D_i = (1 - d), \mathbf{X}_i = \mathbf{x}] \quad (d \in \{0, 1\})$$

が成り立つことを利用すれば, 最後の等式は正当化できる:

$$\begin{aligned}\mathbb{E}[Y_i(1) - Y_i(0) \mid D_i = 1, \mathbf{X}_i] =& \mathbb{E}[Y_i(1) \mid D_i = 1, \mathbf{X}_i] - \mathbb{E}[Y_i(0) \mid D_i = 1, \mathbf{X}_i] \\ \overset{\text{非交絡}}{=}& \mathbb{E}[Y_i(1) \mid \mathbf{X}_i] - \mathbb{E}[Y_i(0) \mid \mathbf{X}_i] = \tau(\mathbf{X}_i)\end{aligned}$$

以上から，ATE, ATET も識別可能な表現で得られた. □

■p.20 **非交絡の仮定の構造モデルにおける意義** **p.3** で述べた通り, 誤差項が加法分離可能な構造モデルについては, 平均独立の仮定を置くことにより識別の条件は満たされていた. ここでは, より一般的に, 誤差項が加法分離不可能な **(1.14) 式** に基づく構造モデルにおける識別の条件は, 先述の通り平均独立の仮定では不十分で, $f(e \mid D, \mathbf{X}) = f(e \mid \mathbf{X})$ であると結論できる.

Theorem: 一般の構造モデル識別に必要な仮定

誤差項 e が加法分離不可能な以下の構造モデルに基づく因果推論を考える:

$$Y = h(D, \mathbf{X}, e)$$

$h(\cdot)$ は e に関して単調. 関心のある因果効果 $h(1, \mathbf{X}, e) - h(0, \mathbf{X}, e)$ は以下の条件で識別される:

$$f(e \mid D, \mathbf{X}) = f(e \mid \mathbf{X})$$

2 変数選択

■p.23 **変数選択の目的** 線形回帰モデルにおける共変量の選択はデータ依存的になり得, そこで活用されるのが **変数選択** である. この手法は予測を目的として開発されるのが主であるため, 本章でも予測の観点から議論が進んでいる. 具体的には, 標本内予測誤差を導入してバイアスと分散のトレードオフの議論がなされる.

■p.23 **過学習や** over fitting 別概念なのだろうか.

■p.23 **問題提起 変数選択を行った後の (限界効果の) 統計的推測は不適切. 2.6 節** で議論される.

2.1 設定

■p.24 **定式化** 回帰モデルの仮定を満たしていることが分かる. 均一分散は簡単化のための仮定.

■p.24 **冗長な (コントロール) 変数** 冗長な変数の問題点を確認しておこう. コントロール変数を不用意に追加することは, 一致性には影響を及ぼさずとも, 推定量の分散を大きくし, 推定精度を悪化させる.

Example: 真の回帰モデル:

$$Y = \beta_0 + \beta_1 X_1 + e$$

を, 余分に説明変数 X_2 を付け加えた以下の回帰モデル:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$$

で誤って推定する場合を考える．設定から $\beta_2 = 0$. 誤ったモデルも回帰モデルなので, OLS 推定量の真のパラメータとの一致性は保たれる．大標本においては:

$$b_1 \xrightarrow{d} \mathcal{N}(\beta_1, \sigma_{b_1}^2) \quad \text{where } \sigma_{b_1}^2 = \frac{1}{n}\left(\frac{1}{1-\rho_{X_1,X_2}^2}\right)\frac{\sigma_e^2}{\sigma_{X_1}^2}$$

であることが知られている．この結果より, 冗長な変数と興味のある説明変数との間に強い相関 ρ_{X_1,X_2} があるほど推定量の分散が大きくなり, 推定精度は悪化することが分かった．仮説検定を念頭に置けば, 分散の拡大は帰無仮説棄却されづらくし, **第二種の過誤** を深刻化させると言えよう．

■p.24 **(2.2) 式 の指すところ**　少なくとも線形射影モデルではある．

■p.25 **標本内予測誤差**　MSE, MSPE とはべつの指標.

$$\mathrm{Err}_{\mathrm{in}}(M) = \mathbb{E}[\frac{1}{n}\|\tilde{\mathbf{Y}} - \mathbf{X}_M\widehat{\boldsymbol{\beta}}_M\|^2 \mid \mathbf{X}, \mathbf{Y}] \tag{2.3}$$

観測される全変数で条件が構成されてることに注意せよ．$\mathrm{ErrS}(M)$ は標本内予測誤差の期待値をとったもの:

$$\mathrm{ErrS}(M) = \mathbb{E}[\mathrm{Err}_{\mathrm{in}}(M)] = \mathbb{E}[\frac{1}{n}\|\tilde{\mathbf{Y}} - \mathbf{X}_M\widehat{\boldsymbol{\beta}}_M\|^2] \tag{2.4}$$

変数 ($\Longleftrightarrow$ 線形モデル M) を選択して上記の二つを最小化するのが目的となる．ここで, $\tilde{\mathbf{Y}}$ は $\mathbf{X}$ を条件として $\mathbf{Y}$ と i.i.d. な確率変数ベクトルである．つまり, $\tilde{\mathbf{Y}}$ は **予測に使用していないデータ** である．

■p.26 **標本内**, in-sample　以下のように定義されている．

Def: 標本内 (in-sample)

標本と同じ共変量 $\mathbf{X}$ を用いて, 未知の $\tilde{\mathbf{Y}}$ を予測すること．

ここで注意すべきは, $\tilde{\mathbf{Y}}$ が既知の $\mathbf{X}$ をそっくりそのまま共変量として持つこと．イメージは以下の図:

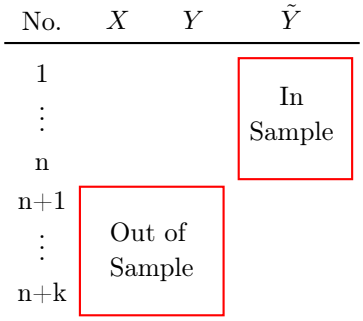

図 3　In-sample と Out-of-sample

■p.26 **期待値の違い**

- **(2.3) 式**: ある標本 $(\mathbf{X}, \mathbf{Y})$ の下でのモデルで推定された予測 $\widehat{\boldsymbol{\mu}}_M(\mathbf{X})$ の平均的な予測誤差
- **(2.4) 式**: 繰り返し標本を得た際に, 同じモデル M を使い続けた時の予測誤差の平均

■p.26 **注 3**　In Sample が Same-X と呼ばれる理由は上記の図に整理した通り, 同じ共変量を用いるため. 妥当性の検証については **2.5 節** を待つ. モデル選択に終始して, 予測をしていない, 若しくは予測を同時に行なっていることが問題.

■p.26 $\mathrm{Err}_{\mathrm{in}}$ **最小化の別解釈**　$\mathrm{Err}_{\mathrm{in}}(M)$ と $L(M) = \frac{1}{n}\|\boldsymbol{\mu} - \mathbf{X}_M\widehat{\boldsymbol{\beta}}_M\|^2$ の minimizer の一致を確認する.

Proof. 以下のように変形することで目標の関係式を得る:

$$\begin{aligned}
\mathrm{Err}_{\mathrm{in}}(M) &= \mathbb{E}[\frac{1}{n}\|(\tilde{\mathbf{Y}} - \boldsymbol{\mu}) + (\boldsymbol{\mu} - \mathbf{X}_M\widehat{\boldsymbol{\beta}}_M)\|^2 \mid \mathbf{X}, \mathbf{Y}] \\
&= \mathbb{E}[\frac{1}{n}\|\tilde{\mathbf{Y}} - \boldsymbol{\mu}\|^2 \mid \mathbf{X}, \mathbf{Y}] + \mathbb{E}[\frac{2}{n}(\tilde{\mathbf{Y}} - \boldsymbol{\mu})^\top(\boldsymbol{\mu} - \mathbf{X}_M\widehat{\boldsymbol{\beta}}_M) \mid \mathbf{X}, \mathbf{Y}] + \frac{1}{n}\|\boldsymbol{\mu} - \mathbf{X}_M\widehat{\boldsymbol{\beta}}_M\|^2 \\
&= \mathbb{E}[\frac{1}{n}\|\tilde{\mathbf{Y}} - \boldsymbol{\mu}\|^2 \mid \mathbf{X}, \mathbf{Y}] + \frac{2}{n}(\boldsymbol{\mu} - \mathbf{X}_M\widehat{\boldsymbol{\beta}}_M)^\top\mathbb{E}[\tilde{\mathbf{Y}} - \boldsymbol{\mu} \mid \mathbf{X}, \mathbf{Y}] + L(M) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(\tilde{Y}_i - \mu(\mathbf{X}_i))^2 \mid \mathbf{X}, \mathbf{Y}] + L(M) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(\tilde{Y}_i - \mathbb{E}[\tilde{Y}_i \mid \mathbf{X}])^2 \mid \mathbf{X}] + L(M) \\
&= \frac{1}{n}\cdot n\cdot\sigma^2 + L(M) = \sigma^2 + L(M)
\end{aligned}$$

ここで σ^2 は条件付分散として定義されており, モデルに依存しないことも確認できる. □

$\tilde{\mathbf{Y}}$ の予測に関心があるときには $\mathrm{Err}_{\mathrm{in}}$ や ErrS に注目する一方で, 回帰関数 $\boldsymbol{\mu}$ の推定に興味がある際には損失やその期待値であるリスクに注目する. これは回帰関数に限定されず一般の回帰モデルに拡張できそうだ.

2.2 推定された予測誤差に基づく変数選択

■p.27 **推定対象**　推定したいのは $\mathrm{ErrS}(M)$ だが, $\tilde{\mathbf{Y}}$ は観測されないため観測可能な推定量を構成する必要がある. この要請から, (ad hoc な) 置き換えによって以下の式は導かれた:

$$\bar{\mathrm{err}}(M) = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}_M\widehat{\boldsymbol{\beta}}_M\|^2 \tag{2.5}$$

この推定量のバイアス (後述) を訂正する手法を探るのがこの節の目標である.

■p.27 **説明変数の数と** (2.5)　記載があるように, **(2.5) 式** は残差二乗和 (SSR) をサンプルサイズで割ったものに等しい. すなわち, **(2.5) 式** の最小化問題は, 決定係数の最大化問題に他ならない. 決定係数は説明変数を増やせば上昇することが知られている[*6] ため, 最大のモデル $\{1, \ldots, p\}$ が選択されることが確認できる.

■p.28 **オーバーフィッティング**　以下のように分解が可能:

$$\mathbb{E}[L(M) \mid \mathbf{X}] = \frac{1}{n}\|\boldsymbol{\mu} - \mathbf{H}_M\boldsymbol{\mu}\|^2 + \frac{\sigma^2 p_M}{n}$$

ここで, $\mathbf{H}_M$ は射影行列で以下のように定義される:

$$\mathbf{H}_M = \mathbf{X}_M(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}\mathbf{X}_M^\top$$

[*6] SSR の最小化問題でパラメータの制約が緩和されることに他ならない.

Proof 射影行列が $\mathbf{X}$ 可測であることを利用すれば以下のように変形が可能.

$$\begin{aligned}\mathbb{E}[L(M) \mid \mathbf{X}] &= \mathbb{E}[\frac{1}{n}\|\boldsymbol{\mu} - \mathbf{X}_M\widehat{\boldsymbol{\beta}}_M\|^2 \mid \mathbf{X}] = \mathbb{E}[\frac{1}{n}\|\boldsymbol{\mu} - \mathbf{X}_M(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}\mathbf{X}_M^\top\mathbf{Y}\|^2 \mid \mathbf{X}] \\ &= \mathbb{E}[\frac{1}{n}\|\boldsymbol{\mu} - \mathbf{H}_M(\boldsymbol{\mu}+\mathbf{e})\|^2 \mid \mathbf{X}] \\ &= \frac{1}{n}\|\boldsymbol{\mu} - \mathbf{H}_M\boldsymbol{\mu}\|^2 + \mathbb{E}[\frac{1}{n}\|\mathbf{H}_M\mathbf{e}\|^2 \mid \mathbf{X}] - \frac{2}{n}\boldsymbol{\mu}^\top\mathbf{H}_M\mathbb{E}[\mathbf{e} \mid \mathbf{X}] \\ &= \frac{1}{n}\|\boldsymbol{\mu} - \mathbf{H}_M\boldsymbol{\mu}\|^2 + \mathbb{E}[\frac{1}{n}\|\mathbf{H}_M\mathbf{e}\|^2 \mid \mathbf{X}]\end{aligned}$$

第二項が目標の表現になることを, 射影行列の冪等性 $\mathbf{H}_M^\top\mathbf{H}_M = \mathbf{H}_M$ を利用して以下のように示す:

$$\begin{aligned}\mathbb{E}[\frac{1}{n}\|\mathbf{H}_M\mathbf{e}\|^2 \mid \mathbf{X}] &= \frac{1}{n}\mathbb{E}[\mathbf{e}^\top\mathbf{H}_M^\top\mathbf{H}_M\mathbf{e} \mid \mathbf{X}] = \frac{1}{n}\mathbb{E}[\mathbf{e}^\top\mathbf{H}_M\mathbf{e} \mid \mathbf{X}] \\ &= \frac{1}{n}\mathrm{tr}(\mathbf{H}_M\Sigma) = \frac{1}{n}\sigma^2\mathrm{tr}(\mathbf{H}_M\mathbf{I}) = \frac{\sigma^2 p_M}{n}\end{aligned}$$

ここでは均一分散, 射影行列のトレースが説明変数の数に一致すること, そして二次形式の期待値とトレースの関係を利用した. □

■導出に利用した関係性　証明しておこう.

Thm: 条件付加重共分散行列とトレース

$\mathbf{e} \in \mathbb{R}^n$ が条件付きで $\mathbb{E}[\mathbf{e} \mid \mathbf{X}] = \mathbf{0}$, $\mathrm{Cov}[\mathbf{e} \mid \mathbf{X}] = \sigma^2\mathbf{I}$ を満たすとき, 任意の $\mathbf{X}$ 可測な対称行列 $\mathbf{A} \in \mathbb{R}^{n\times n}$ に対して次が成り立つ:

$$\mathbb{E}[\mathbf{e}^\top\mathbf{A}\mathbf{e} \mid \mathbf{X}] = \sigma^2\mathrm{tr}(\mathbf{A})$$

Proof $\mathbf{e}^\top\mathbf{A}\mathbf{e}$ はスカラーで, トレースの可換不変性 から以下のように変形できる:

$$\mathbb{E}[\mathbf{e}^\top\mathbf{A}\mathbf{e} \mid \mathbf{X}] = \mathbb{E}\left[\mathrm{tr}(\mathbf{e}^\top\mathbf{A}\mathbf{e}) \mid \mathbf{X}\right] = \mathbb{E}\left[\mathrm{tr}(\mathbf{A}\mathbf{e}\mathbf{e}^\top) \mid \mathbf{X}\right] = \mathrm{tr}\left(\mathbf{A}\cdot\mathbb{E}[\mathbf{e}\mathbf{e}^\top \mid \mathbf{X}]\right)$$

ここで $\mathbb{E}[\mathbf{e} \mid \mathbf{X}] = \mathbf{0}$ より, $\mathbb{E}[\mathbf{e}\mathbf{e}^\top \mid \mathbf{X}] = \mathrm{Cov}[\mathbf{e} \mid \mathbf{X}] = \sigma^2\mathbf{I}$. したがって:

$$\mathbb{E}[\mathbf{e}^\top\mathbf{A}\mathbf{e} \mid \mathbf{X}] = \mathrm{tr}(\mathbf{A}\cdot\sigma^2\mathbf{I}) = \sigma^2\mathrm{tr}(\mathbf{A})$$

であり, 題意は満たされた. □

Thm: 射影係数のトレースと説明変数の数

射影行列のトレースとランクは, モデルの説明変数の数に一致する.

Proof $\mathbf{X} \in \mathbb{R}^{n\times k}$ を列フルランク（$\mathrm{rank}(\mathbf{X}) = k$）の行列とする. 示したいのは $\mathrm{tr}(\mathbf{H}) = k = p_M$ である. $\mathbf{H}$ は対称行列より, 直交行列 $\mathbf{Q}$ と対角行列 Λ によって以下のように直交対角化が可能

$$\mathbf{Q}^\top\mathbf{H}\mathbf{Q} = \Lambda,\ \mathbf{H} = \mathbf{Q}\Lambda\mathbf{Q}^\top,\quad \text{where } \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n),\quad \mathbf{Q}^\top\mathbf{Q} = \mathbf{Q}\mathbf{Q}^\top = \mathbf{I}.$$

ここで, $\boldsymbol{\lambda}$ は射影行列の固有値である. 冪等性 $\mathbf{H}^2 = \mathbf{H}$ より:

$$\mathbf{Q}\Lambda\mathbf{Q}^\top = \mathbf{Q}\Lambda\mathbf{Q}^\top\mathbf{Q}\Lambda\mathbf{Q}^\top = \mathbf{Q}\Lambda^2\mathbf{Q}^\top \implies \Lambda^2 = \Lambda \implies \lambda_i^2 = \lambda_i \quad (\forall i = 1, \ldots, n),$$

すなわち各固有値 λ_i は 0 または 1 である. トレースの可換不変性 より:

$$\mathrm{tr}(\mathbf{H}) = \mathrm{tr}(\mathbf{Q}\Lambda\mathbf{Q}^\top) = \mathrm{tr}(\Lambda\mathbf{Q}\mathbf{Q}^\top) = \mathrm{tr}(\Lambda) = \sum_{i=1}^{n} \lambda_i = \#\{\, i \mid \lambda_i = 1\}.$$

一方, ランクは非零固有値の個数に等しいので

$$\mathrm{rank}(\mathbf{H}) = \#\{\, i \mid \lambda_i \neq 0\} = \#\{\, i \mid \lambda_i = 1\}.$$

また $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ で フルランク行列の積とランクの関係, 識別のために置いた $(\mathbf{X}^\top\mathbf{X})^{-1}$ の正則性の仮定 から, $\mathrm{rank}(\mathbf{H}) = \mathrm{rank}(\mathbf{X}) = k$ が成り立つ. よって:

$$\mathrm{tr}(\mathbf{H}) = \#\{\, i \mid \lambda_i = 1\} = \mathrm{rank}(\mathbf{H}) = k.$$

が成り立つことが分かった. □

■p.28 **分解の意味**　さて, 導出が出来たところで以下の分解の意味を精査しよう:

$$\mathbb{E}[L(M) \mid \mathbf{X}] = \frac{1}{n}\|\boldsymbol{\mu} - \mathbf{H}_M\boldsymbol{\mu}\|^2 + \frac{\sigma^2 p_M}{n}$$

第一項のノルムの中は, $\boldsymbol{\mu}$ を $\mathbf{X}_M$ の張る射影空間に写した射影残差を表しており, $\boldsymbol{\mu}$ が射影空間上に既にある場合, 即ち正しく定式化されている場合は, 自分自身に射影される ため, この項は 0 となる. なお, この議論は損失を利用していることから, 回帰関数の推定を念頭に置いていることに注意せよ. 第二項は推定量の分散を表しており, p_M の増加関数であることから, 変数の追加に対する罰則として解釈できる.

■p.28 ad hoc **な推定量の問題点**　**(2.5) 式** はバイアスを持つため修正の必要がある. これを確認しよう.

Proof $L(M) = \frac{1}{n}\|\boldsymbol{\mu} - \mathbf{X}_M\widehat{\boldsymbol{\beta}}_M\|^2$ である. $\mathbf{Y} = \boldsymbol{\mu} + \mathbf{e}$ を用いて, $\bar{\mathrm{err}}(M)$ を展開する：

$$\begin{aligned}
\bar{\mathrm{err}}(M) &= \frac{1}{n}\|\boldsymbol{\mu} + \mathbf{e} - \mathbf{X}_M\widehat{\boldsymbol{\beta}}_M\|^2 = \frac{1}{n}\|(\boldsymbol{\mu} - \mathbf{X}_M\widehat{\boldsymbol{\beta}}_M) + \mathbf{e}\|^2 \\
&= L(M) + \frac{1}{n}\|\mathbf{e}\|^2 + \frac{2}{n}\mathbf{e}^\top\boldsymbol{\mu} - \frac{2}{n}\mathbf{e}^\top\mathbf{X}_M\widehat{\boldsymbol{\beta}}_M \qquad (2.6) \\
&= L(M) + \frac{1}{n}\|\mathbf{e}\|^2 + \frac{2}{n}\mathbf{e}^\top\boldsymbol{\mu} - \frac{2}{n}\mathbf{e}^\top\mathbf{H}_M\mathbf{Y}
\end{aligned}$$

期待値をとると, 第 3 項は $\mathbb{E}[\mathbb{E}[\mathbf{e} \mid \mathbf{X}]] = \mathbb{E}[\mathbf{e}] = \mathbf{0}$ より消えるため:

$$\begin{aligned}
\mathbb{E}[\bar{\mathrm{err}}(M)] &= \mathbb{E}[L(M)] + \frac{1}{n}\mathbb{E}[\mathbb{E}[\mathbf{e}^\top\mathbf{e} \mid \mathbf{X}]] - \frac{2}{n}\mathbb{E}[\mathbf{e}^\top\mathbf{X}_M\widehat{\boldsymbol{\beta}}_M] \\
&= \mathbb{E}[L(M)] + \frac{1}{n}\mathbb{E}[\mathrm{tr}(\mathrm{Cov}[\mathbf{e} \mid \mathbf{X}])] - \frac{2}{n}\mathbb{E}[\mathbf{e}^\top\mathbf{X}_M\widehat{\boldsymbol{\beta}}_M] \\
&= \mathbb{E}[L(M)] + \frac{1}{n}\cdot n\cdot\sigma^2 - \frac{2}{n}\mathbb{E}[\mathbf{e}^\top\mathbf{X}_M\widehat{\boldsymbol{\beta}}_M] = \mathbb{E}[L(M) + \sigma^2] - \frac{2}{n}\mathbb{E}[\mathbf{e}^\top\mathbf{X}_M\widehat{\boldsymbol{\beta}}_M] \\
&= \mathbb{E}[\mathrm{Err}_{\mathrm{in}}(M)] - \frac{2}{n}\mathbb{E}[\mathbf{e}^\top\mathbf{X}_M\widehat{\boldsymbol{\beta}}_M] = \mathrm{ErrS}(M) - \frac{2}{n}\mathbb{E}[\mathbf{e}^\top\mathbf{X}_M\widehat{\boldsymbol{\beta}}_M]
\end{aligned}$$

ここでは $\mathrm{Err}_{\mathrm{in}}(M) = L(M) + \sigma^2$ を利用した. 更に, $\mathbb{E}[\mathbf{e}^\top \mathbf{H}_M \mathbf{e}] = \sigma^2 p_M$ により, 以下が成り立つ:

$$\mathbb{E}[\mathbf{e}^\top \mathbf{X}_M \boldsymbol{\beta}_M] = \mathbb{E}[\mathbf{e}^\top \mathbf{H}_M \mathbf{Y}] = \mathbb{E}[\mathbb{E}[\mathbf{e}^\top \mid \mathbf{X}]\mathbf{H}_M \boldsymbol{\mu}] + \mathbb{E}[\mathbf{e}^\top \mathbf{H}_M \mathbf{e}] = \sigma^2 p_M$$

これを代入して:

$$\mathbb{E}[\bar{\mathrm{err}}(M)] = \mathrm{ErrS}(M) - \underbrace{\frac{2\sigma^2 p_M}{n}}_{\text{バイアス, オプティミズム}}$$

を得る. 過少推定されており, 説明変数の増加がこれを深刻化させることも分かった. □

■p.29 Mallows の C_p 基準 (Mallows, 1973)　オプティミズムの分だけ上方修正した推定量を利用する:

$$C_p(M) = \bar{\mathrm{err}}(M) + \frac{2\sigma^2 p_M}{n}$$

当然不偏性が回復する. σ^2 は未知であるからこれを推定する必要がある.

■p.29 Efron, 2004; Rosset and Tibshirani, 2020　オプティミズムは線形回帰以外にも概念を拡張可能である.

Proof 先ず, 線形推定を念頭に置いたオプティミズムは以下のように書けた:

$$\mathrm{Opt} \coloneqq \mathbb{E}[\mathrm{Err}_{\mathrm{in}}(M) - \bar{\mathrm{err}}(M)] = \mathbb{E}[\frac{1}{n}\|\tilde{\mathbf{Y}} - \mathbf{X}_M \widehat{\boldsymbol{\beta}}_M\|^2 - \frac{1}{n}\|\mathbf{Y} - \mathbf{X}_M \widehat{\boldsymbol{\beta}}_M\|^2]$$

推定の部分を $\widehat{\boldsymbol{\mu}}$ に一般化することでオプティミズムも一般化できる:

$$\mathrm{Opt}(M) \coloneqq \mathbb{E}[\frac{1}{n}\|\tilde{\mathbf{Y}} - \widehat{\boldsymbol{\mu}}(\mathbf{X})\|^2 - \frac{1}{n}\|\mathbf{Y} - \widehat{\boldsymbol{\mu}}(\mathbf{X})\|^2]$$

LIE を利用すれば, $\tilde{\mathbf{Y}}^2$ と $\mathbf{Y}^2$ は $\mathbf{X}$ の条件付きで i.i.d. なため打ち消せ, 残るのは交差項のみ:

$$\begin{aligned}
\mathrm{Opt} \coloneqq& \mathbb{E}[\mathbb{E}[-\frac{2}{n}\tilde{\mathbf{Y}}^\top \widehat{\boldsymbol{\mu}}(\mathbf{X}) + \frac{2}{n}\mathbf{Y}^\top \widehat{\boldsymbol{\mu}}(\mathbf{X}) \mid \mathbf{X}]] = \frac{2}{n}\mathbb{E}[\mathbb{E}[\mathbf{Y}^\top \widehat{\boldsymbol{\mu}}(\mathbf{X}) - \tilde{\mathbf{Y}}^\top \widehat{\boldsymbol{\mu}}(\mathbf{X}) \mid \mathbf{X}]] \\
=& \frac{2}{n}\mathbb{E}[\mathbb{E}[\mathbf{Y}^\top \widehat{\boldsymbol{\mu}}(\mathbf{X})) \mid \mathbf{X}] - \mathbb{E}[\tilde{\mathbf{Y}}^\top \mid \mathbf{X}]\mathbb{E}[\widehat{\boldsymbol{\mu}}(\mathbf{X}) \mid \mathbf{X}]] \\
=& \frac{2}{n}\mathbb{E}[\mathbb{E}[\mathbf{Y}^\top \widehat{\boldsymbol{\mu}}(\mathbf{X})) \mid \mathbf{X}] - \mathbb{E}[\mathbf{Y}^\top \mid \mathbf{X}]\mathbb{E}[\widehat{\boldsymbol{\mu}}(\mathbf{X}) \mid \mathbf{X}]] \\
=& \frac{2}{n}\mathbb{E}[\mathrm{tr}(\mathrm{Cov}(\mathbf{Y}, \widehat{\boldsymbol{\mu}}(\mathbf{X}) \mid \mathbf{X}))] = \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}[\,\mathrm{Cov}(Y_i, \widehat{\mu}(\mathbf{X}_i) \mid \mathbf{X})] \qquad (2.7)
\end{aligned}$$

Y_i, $\mu(\mathbf{X}_j)$ $i \neq j$ が独立であることにより, 最後の行の等式が成り立つ. □

■p.30 **共分散罰則**　共分散がフィットとともに高くなることによってトレードオフが起こる. 文脈からして $+$ の共分散を持つが, これが実際に成立することを, 線形回帰の場合に限ってだが確認しておこう.

Proof 先ほど得たオプティミズムの共分散での表現を線形回帰に特定化してみよう:

$$\begin{aligned}
\mathrm{Opt} =& \frac{2}{n}\mathbb{E}[\mathrm{tr}(\mathrm{Cov}(\mathbf{Y}, \widehat{\boldsymbol{\mu}}_M(\mathbf{X}) \mid \mathbf{X}))] = \frac{2}{n}\mathbb{E}[\mathrm{tr}(\mathrm{Cov}(\mathbf{Y}, \mathbf{H}_M \mathbf{Y} \mid \mathbf{X}))] \\
=& \frac{2}{n}\mathbb{E}[\mathrm{tr}(\mathbf{H}_M \mathrm{Cov}(\mathbf{Y}, \mathbf{Y} \mid \mathbf{X}))] = \frac{2}{n}\mathrm{tr}(\mathbf{H}_M \mathbf{I}\sigma^2) = \frac{2\sigma^2}{n}\mathrm{tr}(\mathbf{H}_M) = \frac{2\sigma^2 p_M}{n} > 0
\end{aligned}$$

線形回帰の場合に限って得たオプティミズムとも一致することが確認できた. また, 分散が厳密に正であり, 説明変数の個数が (定数回帰も一つとして数えて) 厳密に正であることがほとんどであるため, 結果的にオプティミズムはほぼ常に厳密に正であることが確認できた。 □

■p.30 Stein **の公式**　認めよう.

■p.30 Stein's unbiased risk estimator (SURE)　標本対応によって得られる.

Proof 示したいのは以下の関係式:

$$\sum_{i=1}^{n} \mathbb{E}[\mathrm{Cov}(Y_i, \widehat{\mu}(\mathbf{X}_i) \mid \mathbf{X})] \overset{\text{標本対応}}{\longrightarrow} \sigma^2 \sum_{i=1}^{n} \frac{\partial \widehat{\mu}(\mathbf{X}_i)}{\partial Y_i}$$

$\mathbf{X}_i$ は非確率的で, Stein の公式から仮定のもとで以下が成り立つ:

$$\mathrm{Cov}(Y_i, \widehat{\mu}(\mathbf{X}_i)) = \sigma^2 \, \mathbb{E}[\frac{\partial \widehat{\mu}(\mathbf{X}_i)}{\partial Y_i}]$$

$\mathbf{X}_i$ が非確率的であることから, $\mathbf{X}_i$ の条件付オペレータはすべて条件の無い形に書き換えられる. しかし, $\mathbf{X}_i \neq \mathbf{X}_j \quad i \neq j$ は依然一般に成り立ち, 期待値の値も i に依存する. 以下のように変形可能:

$$\begin{aligned}\sum_{i=1}^{n} \mathbb{E}[\mathrm{Cov}(Y_i, \widehat{\mu}(\mathbf{X}_i) \mid \mathbf{X})] \quad &= \quad \sum_{i=1}^{n} \mathbb{E}[\mathrm{Cov}(Y_i, \widehat{\mu}(\mathbf{X}_i))] \overset{\text{LIE}}{=} \sum_{i=1}^{n} \sigma^2 \, \mathbb{E}[\frac{\partial \widehat{\mu}(\mathbf{X}_i)}{\partial Y_i}] \\ &\overset{\text{標本対応}}{\longrightarrow} \frac{\sigma^2}{n} \sum_{i=1}^{n} \underbrace{\sum_{i=1}^{n} \frac{\partial \widehat{\mu}(\mathbf{X}_i)}{\partial Y_i}}_{i \text{ に依存しない}} = \frac{n\sigma^2}{n} \sum_{i=1}^{n} \frac{\partial \widehat{\mu}(\mathbf{X}_i)}{\partial Y_i} = \sigma^2 \sum_{i=1}^{n} \frac{\partial \widehat{\mu}(\mathbf{X}_i)}{\partial Y_i}\end{aligned}$$

よって目標の関係式が示された. □

■p.31 **一個抜き交差検証法の性質**　i のデータは独立なので, 条件付期待値を駆使して以下の変形が可能:

$$\mathbb{E}[\sum_{i=1}^{n} e_i \mathbf{X}_{M,i}^{\top} \widehat{\boldsymbol{\beta}}_{M,(-i)}] = \sum_{i=1}^{n} \mathbb{E}[\mathbb{E}[e_i \mid \mathbf{X}, \mathbf{Y}_{(-i)}] \, \mathbf{X}_{M,i}^{\top} \widehat{\boldsymbol{\beta}}_{M,(-i)}] = \sum_{i=1}^{n} \mathbb{E}[\mathbb{E}[e_i \mid \mathbf{X}] \, \mathbf{X}_{M,i}^{\top} \widehat{\boldsymbol{\beta}}_{M,(-i)}] = 0$$

この結果を利用して, 期待値をとれば第一, 二項だけが残るために以下のように結論付けられる:

$$\mathbb{E}[\mathrm{CV}_{\mathrm{LOO}}(M)] \approx \mathbb{E}[L(M) + \sigma^2] = \mathrm{ErrS}(M)$$

近似になっているのは $\widehat{\boldsymbol{\beta}}_{M,(-i)}$ が $\widehat{\boldsymbol{\beta}}_M$ を代替するためで, この差がちいさければ, つまり leverage が小さければ近似は正統化できる. また, モデル選択のためには回帰を k 回繰り返す必要があるが, この計算負荷軽減のために後述の leverage が利用される.

■p.32 leverage **の性質**　**Hansen (2022)** の **Ch 3.19∼21** に説明がある. ここではこれを確認してみよう.

Thm 3.6 (Hansen): Properties of Leverage

1. $0 \leq h_{ii} \leq 1$
2. $h_{ii} \geq 1/n$ if X includes an intercept.
3. $\sum_{i=1}^{n} h_{ii} = k$

Proof 先ず Hansen に則って一つ目を証明しよう. s_i を $n \times 1$ の単位ベクトルとして, leverage は:

$$h_{ii} = s_i^\top \mathbf{P} s_i$$

Quadratic Inequality と, 射影行列の固有値が k 個の 1 と $n-k$ 個の 0 であることを利用し:

$$h_{ii} = s_i^\top \mathbf{P} s_i \leq \|\mathbf{P}\| s_i^\top s_i \leq 1 \cdot s_i^\top s_i = 1$$

二つ目は Hansen を見よ. 三つ目は 射影行列のトレースが説明変数の数に一致する ことから成立. □

証明に利用したのは以下の不等式.

Thm B.18 (Hansen): Quadratic Inequality

任意の $m \times 1$ ベクトル $\mathbf{b}$ と $m \times m$ 対称行列 $\mathbf{A}$ について以下が成立する:

$$\mathbf{b}^\top \mathbf{A} \mathbf{b} \leq \|\mathbf{A}\| \mathbf{b}^\top \mathbf{b}$$

ここで $\|\mathbf{A}\|$ は 任意の実数行列で定義される $\mathbf{A}$ の spectral norm [*7] で, 最大の singular value である:

$$\|\mathbf{A}\| = \sigma_{\max}(\mathbf{A}) = (\lambda_{\max}(\mathbf{A}^\top \mathbf{A}))^{1/2}$$

$\lambda_{\max}(\mathbf{B})$ は行列 $\mathbf{B}$ の最大の固有値を表す. 行列 $\mathbf{A}$ が対称行列なら, 以下の書き換えが可能である:

$$\|\mathbf{A}\| = \max_{j \leq n} |\lambda_j|$$

■p.32 leverage の意味 h_{ii} は i 番目の観測値 X_i の相対的な異質性を表し, これを測る指標が以下の式:

$$\bar{h} = \max_{1 \leq i \leq n} h_{ii}$$

この値が大きいことを **unbalanced な回帰設計** と言い, 極端な例としては一人だけが当てはまるようなダミー変数回帰が挙げられる. この場合はその人物に最大の leverage として 1 が割り振られる. 計算は 末石 ミクロ計量分析 第二回 を参照すること.

■leverage を用いた LOO の証明 Hansen の **Ch 3.20** を参照しよう. (未確認)

■LOO は標本内か? 標本内の定義からして, Y_i と結びつく $\mathbf{X}_i$ と同様の共変量を持った Y_j $i \neq j$ がいなければ, 標本内推定とは言えない. ゆえに **LOO は標本外推定** であると結論づけられる (**2.5 節**).

2.3 情報量規準

■p.34 最尤推定法 末石 (2014) の **付録 C** を参考に復習しておこう. 誤差項の正規性の仮定のもとでの議論であることは留意しておこう. (未確認)

■p.36 AIC で選択されるモデル C_p 基準と似たモデルが選択される傾向にある. BIC は, $\ln(n) > 2 \implies n \geq 8$ の時に罰則が大きいため, AIC よりも共変量の少ないモデルを推定しがちである.

[*7] Hansen の Appendix A.23: Matrix Norms を参照すること.

2.4 変数選択の一致性と漸近最適性

■p.37 **変数選択の目的**　予測に加え, 最も簡潔な正しいモデルを選択するためにも変数選択は利用される. **後者においては BIC が, 前者には他の手法が適している.** この節ではこれを結論に持つ議論が展開される.

■p.37 **モデル選択の一致性**　以下のように定義されている.

Def: モデル選択の一致性

線形回帰モデルが正しく定式化されていることを前提として, 最も簡潔な正しいモデルを $M_0 \in \mathcal{M}$ と表し, ある変数選択の手法で選択されるモデルを $\widehat{M}$ と記載するならば, 一定の条件下で $n \to \infty$ のとき:

$$\mathbf{P}\left[\widehat{M} = M_0\right] \to 1$$

が成り立つならば, 変数選択・モデル選択に一致性があるという.

BIC は一致性を持つが, AIC は冗長な変数を選びうる.

■p.38 **モデル選択の漸近最適性**　こちらは回帰関数の推定精度測定の目的での有用性をはかる指標.

Def: モデル選択の漸近最適性

ある変数選択の手法で選択されるモデルを $\widehat{M} \in \mathcal{M}$ と記載するならば, 一定の条件下で $n \to \infty$ のとき:

$$\frac{L(\widehat{M})}{\min_{M \in \mathcal{M}} L(M)} \xrightarrow{p} 1$$

が成り立つならば, 変数選択・モデル選択に漸近最適性があるという.

$L(M)$ を $\mathrm{Err}_{\mathrm{in}}(M)$ に置き換えれば, minimizer が同一で結果は変わらないため, 予測精度の基準としても利用できることが分かる. ここでは必ずしも線形回帰モデルは正しく定式化されている必要はない. 一致性との状況とは逆に, BIC は漸近最適性を持たないが, 他の手法はこれを満たす.

■p.39 **漸近最適性と回帰関数の形状**　恐らく一様妥当性を念頭に置いた記述.

2.5 その他のモデル評価基準

■p.40 **LOO の有用性**　先に述べた通り, LOO は標本外推定と言えるため, 標本外予測の性能を示す汎化誤差に対応する $\mathrm{ErrR}(M)$ についてもほぼ不偏な推定量を得られる.

2.6 変数選択後の統計的推測の問題

■p.40 **変数選択と推測**　変数選択は不確実性をその過程に導入する一方で, 統計的推測では正しいモデルをあらかじめ所与のものと見なして推測を行う. このモデルの扱いの差異によって, 限界効果に興味がある時に変数選択を行ってしまうと被覆確率に不具合が生じる. 従って統計的推測を行う際には一先ずの対処法として全ての共変量を入れることになり, 冗長な変数の発生により第二種の過誤は深刻になる. $n < k$ での対処法は **7 章** で扱われる. マクロの時系列では変数選択を積極的に用いるが, これは予測を念頭に置いているため.

3 ノンパラメトリック回帰

3.1 カーネル推定

■p.48 **推定量とカーネル** バンド幅の選択に比べると, カーネルの選択は推定量の性質にそれほど影響を与えないとされている, らしい. u はバンド幅の選択によって変化することに注意せよ.

■p.48 NW **推定量** 最小化問題との対応は明らか. **(3.3) 式** で出る NW 推定量は局所定数推定量とも呼び, これを一般化した **(3.4) 式** は局所線形 (LL) 推定量を与える.

■p.49 LL **推定量** 最小化問題 **(3.5) 式** の minimizer は $\mu(x), \mu'(x)$ の推定量. 局所多項式もテイラー展開に従った (係数) $\times\, \mu^{(n)}(x)$ の推定量が minimizer として導かれる.

■p.50 **推定精度の指標** バンド幅 h を用いた $\mu(x)$ の推定量を $\widehat{\mu}_h(x)$ として, この (条件付) MSE は:

$$\begin{aligned}\mathrm{MSE}(h) &= \mathbb{E}[(\widehat{\mu}_h(x) - \mu(x))^2 \mid \mathbf{X}] \\ &= \mathbb{E}[\widehat{\mu}_h(x) - \mu(x) \mid \mathbf{X}]^2 + \mathbb{E}[(\widehat{\mu}_h(x) - \mathbb{E}[\widehat{\mu}_h(x) \mid \mathbf{X}])^2 \mid \mathbf{X}]\end{aligned}$$

Proof 常套手段で証明可能.

$$\begin{aligned}\mathrm{MSE}(h) &= \mathbb{E}\Big[\big(\widehat{\mu}_h(x) - \mathbb{E}[\widehat{\mu}_h(x) \mid \mathbf{X}] + \mathbb{E}[\widehat{\mu}_h(x) \mid \mathbf{X}] - \mu(x)\big)^2 \mid \mathbf{X}\Big] \\ &= \mathbb{E}\Big[\big(\widehat{\mu}_h(x) - \mathbb{E}[\widehat{\mu}_h(x) \mid \mathbf{X}]\big)^2 \mid \mathbf{X}\Big] + \big(\mathbb{E}[\widehat{\mu}_h(x) \mid \mathbf{X}] - \mu(x)\big)^2 \\ &\quad + 2\mathbb{E}\Big[\big(\widehat{\mu}_h(x) - \mathbb{E}[\widehat{\mu}_h(x) \mid \mathbf{X}]\big)\big(\mathbb{E}[\widehat{\mu}_h(x) \mid \mathbf{X}] - \mu(x)\big) \mid \mathbf{X}\Big] \\ &= \mathbb{E}\Big[\big(\widehat{\mu}_h(x) - \mathbb{E}[\widehat{\mu}_h(x) \mid \mathbf{X}]\big)^2 \mid \mathbf{X}\Big] + \big(\mathbb{E}[\widehat{\mu}_h(x) \mid \mathbf{X}] - \mu(x)\big)^2\end{aligned}$$

これは目標の表現である. □

これはバイアスと分散のバンド幅についてのトレードオフの関係を示す. $h \to \infty$ なら一般的なカーネルの形状ならば LL 推定量は通常の線形回帰に帰着することからも, h を大きくすると, 分散は小さく, バイアスは大きくなることが分かる.

■p.50 **近似的な** MSE 漸近的な期待値を求める, ということ? サンプルサイズが大きくなるにつれ, バンド幅を小さくするような設定 ($h \to 0$, $nh^3 \to \infty$) でバイアスを 0 に収束させることが出来るらしい.

■p.51 **漸近的なバイアス** supp を考えるのは, バンド幅 $[x-h, x+h]$ にデータが生成されないことを防ぐため, 生成過程の仮定を弱めるためかも. x の近傍で (高階) 連続微分可能なことを要求するのはテイラー展開に必要だから. また, (有限回) テイラー展開を利用する時点で導関数の存在は仮定されている. q 次の局所多項式推定量では $\mu(x)$ が C^{q+1} 級との仮定は, テイラー展開で必要となるため.

Proof ChatGPT ぱっと出しの証明, 正確性は担保しません. **(3.6) 式** のみ証明しよう.

(1) 条件付き期待値の書き換え

$$\mathbb{E}[\widehat{\mu}_h^{\mathrm{NW}}(x) \mid \mathbf{X}] = \frac{\mathbb{E}\big[K_h(x-X)\,Y \mid \mathbf{X}\big]}{\mathbb{E}\big[K_h(x-X) \mid \mathbf{X}\big]} = \frac{\mathbb{E}\big[K_h(x-X)\,\mu(X)\big]}{\mathbb{E}\big[K_h(x-X)\big]},$$

ここで $K_h(u) = h^{-1}K(u/h)$ とする。経験和を期待値（積分）で近似すると，

$$\mathbb{E}\big[K_h(x-X)\,\mu(X)\big] \approx \int K_h(x-t)\,\mu(t)\,f(t)\,dt.$$

変数変換 $u = (x-t)/h$ をすると

$$\int K_h(x-t)\,\mu(t) f(t)\,dt = \int K(u)\,\mu(x-uh)\,f(x-uh)\,du.$$

同様に分母は

$$\int K(u)\,f(x-uh)\,du.$$

(2) 被積分関数を h について **テイラー展開** する ：

$$\begin{aligned}\mu(x-uh)f(x-uh) = \mu(x)f(x) - (uh)\big(\mu' f + \mu f'\big)(x) \\ + \frac{(uh)^2}{2}\big(\mu'' f + 2\mu' f' + \mu f''\big)(x) + o(h^2),\end{aligned}$$

および

$$f(x-uh) = f(x) - uh\,f'(x) + \frac{(uh)^2}{2} f''(x) + o(h^2).$$

カーネルが対称なら奇次数項（$\int uK(u)\,du$ に乗る項）は消えるので，積分後

$$\int K(u)\mu(x-uh)f(x-uh)\,du = \mu(x)f(x) + \frac{h^2}{2}\kappa_2(K)\big(\mu'' f + 2\mu' f' + \mu f''\big)(x) + o(h^2),$$

$$\int K(u)f(x-uh)\,du = f(x) + \frac{h^2}{2}\kappa_2(K) f''(x) + o(h^2).$$

(3) 比の展開： $A = \mu f$, $a = \mu'' f + 2\mu' f' + \mu f''$, $B = f$, $b = f''$ として

$$\frac{A+\varepsilon a}{B+\varepsilon b} = \frac{A}{B} + \varepsilon\frac{aB - Ab}{B^2} + o(\varepsilon), \qquad \varepsilon = \frac{h^2}{2}\kappa_2(K).$$

ここに代入すると

$$\mathbb{E}[\widehat{\mu}_h^{\mathrm{NW}}(x) \mid \mathbf{X}] = \mu(x) + \frac{h^2}{2}\kappa_2(K)\Big\{\mu''(x) + 2\mu'(x)\frac{f'(x)}{f(x)}\Big\} + o(h^2).$$

したがってバイアス（期待値 − $\mu(x)$）は

$$\mathbb{E}[\widehat{\mu}_h^{\mathrm{NW}}(x) - \mu(x) \mid \mathbf{X}] = \frac{h^2}{2}\kappa_2(K)\mu'\prime(x) + h^2\kappa_2(K)\,\mu'(x)\frac{f'(x)}{f(x)} + o(h^2),$$

これは与式 (3.6) と一致する。

(3.6) 式, **(3.7) 式** より, NW 推定量と LL 推定量のバイアスはおよそ h^2 に比例することが分かる．これをバイアスのオーダーは h^2 ($O(h^2)$) であるという．LL 推定量の方がバイアスは一般的に小さい．

■p.52 LL **推定量の性質, 近似的な** MSE　LL 推定量は漸近的には $(n \to \infty)$, 分散とバイアスを同時に減少させることが可能, つまり**トレードオフを (漸近的に) 乗り越えることが可能である**. 他の局所多項式推定量ではない特異な性質らしい. MSE にバイアスの二乗と分散を代入し, $o_p(\cdot)$ を無視して **(3.8) 式** は得られる. 構成の仕方からも明らかなように, AMSE は漸近的な表現である.

■p.53 MISE　MSE は特定の点における精度評価指標である, **p.6** で述べられていたことだ. サポート全体でいい推定量を判断するためには, 平均積分二乗誤差 (MISE) を用いる:

$$\mathrm{MISE}(h) = \int_{-\infty}^{\infty} \mathbb{E}[(\widehat{\mu}_h(x) - \mu(x))^2 \mid \mathbf{X}] f(x) dx$$

■p.53 **プラグイン法**　AMSE を利用して AMSIE は得られ, MISE の意味で最良のバンド幅は求まる:

$$h_{\mathrm{OPT}} = \left(\frac{R(k) \int \sigma^2(x) dx}{\kappa_2(k)^2 \int (\mu''(x))^2 f(x) dx}\right) n^{-1/5} \tag{3.9}$$

しかし積分部分が未知であるため実行不可能. これを推定したものに置き換えるのが **プラグイン法** である. $\widehat{\mu}$ を推測するために μ'' を仮定する必要があるの?

■p.53 **バンド幅と** LOO　**Ch.2** で ErrS との関係で確認したのと同様の関係:

$$\mathbb{E}[\mathrm{CV}_{\mathrm{LOO}}(h)] = \mathrm{MISE}_{n-1}(h) + \sigma^2$$

が成り立ち, h に依存しない第二項は無視できるため, $\mathbb{E}[\mathrm{CV}_{\mathrm{LOO}}(h)]$ を最小化する h が最適なバンド幅.

■p.54 **次元の呪い**　回帰変数の次元が大きくなるにつれて推定精度が下がる問題のこと. 近傍の観測値が少なくなることで分散が大きくなるために引き起こされる. AMSE を見ても分かるように, 第二項の分散は p で増加する. また, バイアスは陽には p に依存していないが, バンド幅が p に依存し, 増加するため, バイアスのオーダーもまた p に依存し, また p に関して増加する. バンド幅のオーダーは一階条件から求まる. AMSIE ではないことに注意, 特定の点を考えれば十分であるため.

■p.55 **サポートの境界上での回帰関数推定**　**p.52** のあやふやさは承知の上とのこと. X の台が閉区間 $[x_L, x_U]$ のとき, $\mu(x_L)$ の推定は片側しかデータを利用できないことによって NW 推定量ではバイアスが生じる. イメージ図で十分理解できる話だろう. さらにバイアスのオーダーが $O(h^2)$ ではなく $O(h)$ であることから ($\because$ **(3.11) 式** の第一項が h に比例している), 収束が遅く, バイアスがバンド幅を狭めても減っていかないことも分かる. 一方 LL 推定量は $O(h^2)$ のままで, バイアスが小さく保たれる点で優れている.

■p.58 **漸近的にも無視できないバイアスの存在**　$\mathcal{B}$ は漸近的にも無視することが出来ないバイアスで, 信頼区間の構成としてこれを無視する **(3.14) 式** を利用することは **甚だ不適切** である (本文の言い方こえ〜). もう少しまともな対処法の一つとしては, 過少平滑化による $nh^5 \to 0$ となるような h の選択 $(h = n^{-1/4})$ をすること. しかし恣意的な選択をせざるを得ず, これもまた不適切. ゆえに, 近年はバイアスを認めた上で, バイアスに頑健な信頼区間を構成することが, カーネル回帰の主要研究テーマである. **Ch 3.3** では, この流れのひとつとして RDD のもとでのバイアスに頑健な処置効果の信頼区間の構成を確認する.

3.2　シリーズ推定

■p.60 **数学的裏付け**　シリーズ法によるノンパラメトリックモデル回帰を行う. 背景は次の数学的定理:

定理: Stone-Weierstrass

閉区間で定義された連続関数は, 多項式関数によって任意の精度で近似できる, すなわち, $\mu(x)$ が閉区間 $[a,b]$ において連続であれば, 任意の $\varepsilon > 0$ について, ある q と $\boldsymbol{\beta}$ が存在して, 以下が成立する:

$$\max_{x\in[a,b]} |\mu(x) - (\beta_0 - \beta_1 x - \cdots - \beta_q x^q)| < \varepsilon$$

回帰関数を多項式関数で**近似** (× 正しく定式化) することで:

$$Y_i \approx \beta_0 + \beta_1 X_i + \cdots + \beta_q X_i^q + e_i$$

が成立する, この右辺は $\boldsymbol{\beta}$ について線形なので, 形式的に OLS で推定できる. ただ **予測にのみ** 利用可能.

定義: シリーズ法

回帰関数を基底関数の線形結合で近似し, その係数を推定することで関数を推定する方法.

■p.60 **シリーズ推定量**　基底関数ベクトル: $\mathbf{q}_K(x) = (q_1(x), \ldots, q_K(x))^\top$, 近似関数: $\mu_k(x) = \mathbf{q}_K(x)^\top \boldsymbol{\beta}_K$.

$$\boldsymbol{\beta}_K = \arg\min_{\mathbf{b}} \mathbb{E}\big[(\mu(X_i) - \mathbf{q}_K(X_i)^\top \mathbf{b})^2\big] = \mathbb{E}\big[\mathbf{q}_K(X_i)\mathbf{q}_K(X_i)^\top\big]^{-1}\mathbb{E}\big[\mathbf{q}_K(X_i)\mu(X_i)\big]$$

Proof 目的関数 $\mathbb{E}\big[(\mu(X_i) - \mathbf{q}_K(X_i)^\top \mathbf{b})^2\big]$ を $\mathbf{b}$ で最小化する点を求める.

$$\mathbb{E}\big[(\mu(X_i) - \mathbf{q}_K(X_i)^\top \mathbf{b})^2\big] = \mathbb{E}\big[\mu(X_i)^2\big] - 2\,\mathbb{E}\big[\mu(X_i)\mathbf{q}_K(X_i)^\top\big]\mathbf{b} + \mathbf{b}^\top \mathbb{E}\big[\mathbf{q}_K(X_i)\mathbf{q}_K(X_i)^\top\big]\mathbf{b}.$$

FOC は:

$$\mathbf{0} = -2\,\mathbb{E}\big[\mathbf{q}_K(X_i)\mu(X_i)\big] + 2\,\mathbb{E}\big[\mathbf{q}_K(X_i)\mathbf{q}_K(X_i)^\top\big]\mathbf{b}.$$

ここで行列 $\mathbb{E}[\mathbf{q}_K(X_i)\mathbf{q}_K(X_i)^\top]$ が正則, つまり正定値行列ならば, 最小化解は一意に定まり:

$$\mathbf{b} = \mathbb{E}\big[\mathbf{q}_K(X_i)\mathbf{q}_K(X_i)^\top\big]^{-1}\mathbb{E}\big[\mathbf{q}_K(X_i)\mu(X_i)\big]$$

となる. □

μ のシリーズ推定量は:

$$\widehat{\mu}_K(x) = \mathbf{q}_K(x)^\top \widehat{\boldsymbol{\beta}}_K$$

で与えられる. カーネル法では各点 x についてそれぞれ回帰関数の値を推定する必要があるが, シリーズ法は $\boldsymbol{\beta}_K$ の推定だけで終わり. **チューニングパラメータ** として, 基底関数の種類とシリーズの長さ K がある.

■p.61 r **次スプライン関数**　基底関数の例. r 次スプラインの基底関数のベクトルと, 近似関数は:

$$\mathbf{q}_K(x) = \big(1, x, \ldots, x^r, (x-\xi_1)_+^r, \ldots, (x-\xi_l)_+^r\big)^\top$$

$$\mu_K(x) = \mathbf{q}_K(x)^\top \boldsymbol{\beta}_K = \beta_0 + \sum_{j=1}^{r} \beta_j x^j + \sum_{k=1}^{l} \beta_{r+k}(x-\xi_k)_+^r \tag{3.15}$$

$(x)_+ = \max\{x, 0\}$, $\boldsymbol{\xi}$ はノットとよばれ, $\xi_1 < \cdots < \xi_l$ を満たす. $\boldsymbol{\xi}$ は X_i の台を分割しており, **(3.15) 式** は $[\xi_j, \xi_{j+1}]$ で別々の r 次多項式を当てはめ, **$\boldsymbol{\xi}$ で $r-1$ 階導関数が連続になるよう滑らかにつないでいる**. 上級計量で言えば, **Ch 6** で登場した, piecewise continuous のための操作を施した後の **(25) 式** に対応する.

Proof $\mu_K(x)$ の $r-1$ 階導関数を計算すると,

$$\begin{aligned}\frac{d^{r-1}\mu_K(x)}{dx^{r-1}} &= (r-1)!\cdot\beta_{r-1} + r!\cdot\beta_r x + \sum_{k=1}^{l}\beta_{r+k}\frac{d^{r-1}}{dx^{r-1}}(x-\xi_k)_+^r \\ &= (r-1)!\cdot\beta_{r-1} + r!\cdot\beta_r x + r!\cdot\sum_{k:\,x\geq\xi_k}\beta_{r+k}(x-\xi_k)\end{aligned}$$

$[\xi_j,\xi_{j+1}]$ ごとに線形関数で, 和の表現より:

$$\lim_{x\uparrow\xi_j}\frac{d^{r-1}\mu_K(x)}{dx^{r-1}} = \lim_{x\downarrow\xi_j}\frac{d^{r-1}\mu_K(x)}{dx^{r-1}} = \frac{d^{r-1}\mu_K(x)}{dx^{r-1}}\,|_{x=\xi_j}$$

ゆえに連続. だが r 階導関数では連続とはならない. なお, 微分自体も $(\infty,\xi_1)\cup(\xi_1,\xi_2)\cup\cdots\cup(\xi_{l-1},\xi_l)\cup(\xi_l,+\infty)\subsetneq\mathbb{R}$ の区間で行われており, 正確には「滑らか」(C^∞) ではない. □

■p.61 **シリーズの長さ** K 推定精度はシリーズの長さ K に依存する. シリーズ検定量の推定精度の評価には, 以下の **積分二乗誤差 (ISE)** を用いられ, バイアスと分散に分解可能:

$$\begin{aligned}\mathrm{ISE}(K) &\coloneqq \int(\widehat{\mu}_K(x)-\mu(x))^2 f(x)dx \\ &= \underbrace{\int(\mu(x)-\mu_K(x))^2 f(x)dx}_{\text{バイアス: } K\text{で減少}, O(K^{-2\alpha})} + \underbrace{(\widehat{\boldsymbol{\beta}}_K-\boldsymbol{\beta}_K)^\top\Big(\int\mathbf{q}_K(x)\mathbf{q}_K(x)^\top f(x)dx\Big)(\widehat{\boldsymbol{\beta}}_K-\boldsymbol{\beta}_K)}_{\text{分散: } K\text{で上昇}, O_p(K/n)}\end{aligned}$$

Proof 真の回帰関数 $\mu(x)=\mathbb{E}[Y\mid X=x]$ のシリーズ推定量と近似関数はそれぞれ:

$$\widehat{\mu}_K(x)=\mathbf{q}_K(x)^\top\widehat{\boldsymbol{\beta}}_K,\qquad \mu_K(x)=\mathbf{q}_K(x)^\top\boldsymbol{\beta}_K.$$

これを代入して,

$$\begin{aligned}\mathrm{ISE}(K) &= \int((\mathbf{q}_K(x)^\top\widehat{\boldsymbol{\beta}}_K-\mu_K(x))+(\mu_K(x)-\mu(x)))^2 f(x)dx \\ &= \int(\mathbf{q}_K(x)^\top(\widehat{\boldsymbol{\beta}}_K-\boldsymbol{\beta}_K)+(\mu_K(x)-\mu(x)))^2 f(x)dx \\ &= \int(\mu_K(x)-\mu(x))^2 f(x)\,dx + \int(\widehat{\boldsymbol{\beta}}_K-\boldsymbol{\beta}_K)^\top\mathbf{q}_K(x)\mathbf{q}_K(x)^\top(\widehat{\boldsymbol{\beta}}_K-\boldsymbol{\beta}_K)f(x)\,dx \\ &\qquad + 2\int\big((\widehat{\boldsymbol{\beta}}_K-\boldsymbol{\beta}_K)^\top\mathbf{q}_K(x)\big)(\mu_K(x)-\mu(x))f(x)\,dx.\end{aligned}$$

回帰関数はパラメータに関して線形であったことから, 線形射影の定義を利用することができ:

$$\boldsymbol{\beta}_K=\arg\min_{\mathbf{b}}\mathbb{E}[(Y-\mathbf{q}_K(x)^\top\mathbf{b})^2]=\arg\min_{\mathbf{b}}\mathbb{E}[(\mu(x)-\mathbf{q}_K(x)^\top\mathbf{b})^2].$$

ここでは線形射影が回帰関数の最良線形近似でもあることを利用した. FOC は,

$$\begin{aligned}\mathbf{0} &= \frac{\partial}{\partial\mathbf{b}}\mathbb{E}[\mu(x)-\mathbf{q}_K(x)^\top\mathbf{b}] = \frac{\partial}{\partial\mathbf{b}}\int(\mu(x)-\mathbf{q}_K(x)^\top\mathbf{b})^2 f(x)\,dx \\ &= 2\int\mathbf{q}_K(x)(\mu(x)-\mathbf{q}_K(x)^\top\mathbf{b})f(x)\,dx = \int\mathbf{q}_K(x)\big(\mu(x)-\mu_K(x)\big)f(x)\,dx\end{aligned}$$

よって交差項が消え, $(\widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K)$ は x に依存しないため,

$$\mathrm{ISE}(K) = \int (\mu_K(x) - \mu(x))^2 f(x)\,dx + (\widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K)^\top \Big(\int \mathbf{q}_K(x)\mathbf{q}_K(x)^\top f(x)\,dx \Big) (\widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K).$$

これで主張が示された. □

■p.62 **オーダー**　LL 推定量と同様の手順で種々のオーダーについては確認できる. $\mathrm{ISE}(K) = O_p(K^{-4+K/n})$ は足し算でもとまる. 適切なチューニングパラメータを選択できれば, カーネル法・シリーズ法どちらも **オーダーの意味では同じ推定精度** が達成される.

■p.62 **交差検証法による K の選択**　leverage の利用による計算量削減が **(2.9) 式** と同様に可能.

■p.63 **バイアスの存在**　適当な状況の下で, 以下が成立する:

$$\frac{\sqrt{n}(\widehat{\mu}_K(x) - \mu_K(x))}{\sqrt{\mathcal{V}_K(x)}} = \frac{\sqrt{n}\big(\widehat{\mu}_K(x) - \mu(x) + \overbrace{(\mu(x) - \mu_K(x))}^{\text{バイアス: } \mathcal{B}}\big)}{\sqrt{\mathcal{V}_K(x)}} \xrightarrow{d} \mathcal{N}(0,1)$$

カーネル法とは異なり, バイアス $\mu(x) - \mu_K(x) \coloneqq \mathcal{B}$ を **明示的に求められず**, 手元にはオーダーの情報しかないことが問題. 教科書に記載されている信頼区間は, K が十分大きくなるように過小平滑化を行って得られる.

3.3 回帰不連続デザイン, RDD

■p.64 **設定**　定数 c と観測可能なスカラ-変数 X_i が存在して, 処置 D_i について:

$$D_i = \mathbb{1}\{X_i \geq c\}$$

が成り立つとする, この時 X_i をスコア, c をカットオフと呼ぶ. 今回は処置がデジタルに切り替わる **sharp RDD** を考慮する. 恐竜本の **Ch 4** で触れられるような **非遵守者** が存在するような場合には処置はカットオフでは完全には決定されず, これを **fuzzy RDD** と呼ぶ. 詳細は恐竜本 **Ch 7** を見よ.

■p.65 **識別のための条件**　sharp RDD では任意の $X_i = x_i$ で $\mathbf{P}\{D_i = 1 \mid \mathbf{X}_i = \mathbf{x}\} = 0, 1$ となってしまい, オーバーラップの条件が満たされず識別が不能. 興味の対象である $\tau_{\mathrm{rdd}} = \mathbb{E}[Y_i(0) - Y_i(0) \mid X_i = c]$ を推定するための仮定を導入したい, まずスコア X_i は **連続確率変数** として:

Assumption: 連続性

1. $\mathbf{E}[Y_i(1) \mid X_i = x]$, $\mathbf{E}[Y_i(0) \mid X_i = x]$ は $x = c$ において連続
2. X_i の密度関数は, c の近傍で正の値を取る

二つ目の条件は条件付期待値の定義に必要. この下で τ_{rdd} は以下のように識別可能.

$$\widehat{\tau}_{\mathrm{rdd}} = \widehat{\mu}_+ - \widehat{\mu}_- = \lim_{x \downarrow c} \mathbb{E}[Y_i(1) \mid X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i(0) \mid X_i = x] \tag{3.16}$$

Proof 連続性を仮定すると, 次が成り立つ：

$$\mathbb{E}[Y_i(1) \mid X_i = c] = \lim_{x \downarrow c} \mathbb{E}[Y_i(1) \mid X_i = x], \quad x \in [c, \infty)$$
$$\mathbb{E}[Y_i(0) \mid X_i = c] = \lim_{x \uparrow c} \mathbb{E}[Y_i(0) \mid X_i = x], \quad x \in (-\infty, c]$$

sharp RDD では処置変数が閾値で決まるため,

$$\lim_{x \downarrow c} \mathbb{E}[Y_i(1) \mid X_i = x] = \lim_{x \downarrow c} \mathbb{E}[Y_i \mid X_i = x], \quad x \in [c, \infty)$$
$$\lim_{x \uparrow c} \mathbb{E}[Y_i(0) \mid X_i = x] = \lim_{x \uparrow c} \mathbb{E}[Y_i \mid X_i = x], \quad x \in (-\infty, c]$$
$$\mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = c] = \lim_{x \downarrow c} \mathbb{E}[Y_i \mid X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i \mid X_i = x].$$

$\mathbb{E}[Y_i \mid X_i = x]$ は任意の x で観測可能で, 分布が定まれば一意に定まるため, τ_{rdd} も識別可能. □

■多項式を用いた推定　カットオフから離れた推定値に大きなウェイトを与える, カーネル法の一般的なものとは逆の重み付けを行うために不適切.

■p.66 カーネル法を用いた推定　カットオフはそれぞれの関数の **サポートの境界** となるため, NW 推定量はそのバイアスの大きさから不適切ゆえ, LL 推定量を利用して推定する. [*8] サポートの境界のみにおけるよい推定量で充分であるため, 目的関数は MSE で十分.

■p.67 信頼区間構成における課題　Ch 3.1.5 と同様にバイアスが発生する.
p.67 Calonico et al. (2014)　バイアス自体をカーネル法によって推定し, バイアス修正とそれに伴い拡大する分散の補正を行う手法. [*9] 恐竜本では **Ch 6** で詳述されている.

$$\frac{\widehat{\tau}^{\text{bc}}_{\text{rdd}} - \tau_{\text{rdd}}}{\sqrt{V(h) + W(h.b)}} \xrightarrow{d} \mathcal{N}(0,1)$$

ここで, $\widehat{\tau}^{\text{bc}}_{\text{rdd}}$ はバイアス修正した推定量. `rdrobust` はこの手法で与えられる信頼区間を出すコマンドであり, バイアスを修正したうえで検定のサイズを適切に制御した手法との意味で, 恐竜本の恐竜本の p.iii での疑問に答えることが出来た. また, LL 推定量によるバイアス修正の結果は, 同じ h_n を用いた LQ 推定量と等価.
p.68 Calonico et al. (2020)　**(3.17) 式** の被覆確率の誤差を小さくするバンド幅 h の選び方を提案している.

■離散変数のスコア　X_i が離散確率変数では, c 近傍に観測値が存在しない場合があるので, $h \to 0$ **が不能**.
p.68 Kolesar and Rothe (2018)　$\mu(x)$ がある関数のクラス $\mathcal{M}$ に属すると仮定した下で,

$$\lim_{n \to \infty} \inf_{\mu \in \mathcal{M}} \mathbf{P}_\mu(\tau_{\text{rdd}} \in \text{CI}) \geq 1 - \alpha \tag{3.19}$$

を満たすような区間 CI を求める方法を考察する. ここで $\mathbf{P}_\mu(\cdot)$ は真の回帰関数が μ であるときの信頼区間の被覆確率を表す. このような性質を満たす信頼区間を **honest な信頼区間** と呼ぶ. n ごとに異なる (最小の被覆確率を与える) μ を選べるため, honest な信頼区間は n が十分大きければ, 真の μ がどのような関数であっても適切な被覆確率を提供する点で, 一様妥当性を保証すると言える. [*10]

[*8] 末石先生のミクロ計量分析 第 12 回 ではより一般的に局所多項式推定量を利用して推定を行っている.
[*9] 分散が拡大することは AMSE における分散のオーダーを見ればわかる.
[*10] lim と inf は交換できないことに注意せよ.

では具体的に信頼区間を求めよう. 次のような最小化問題を解く:

$$\min_{\alpha,\tau_h,\beta,\gamma} \sum_{i=1}^{n} \mathbb{1}\{|X_i - c| \le h\}\big(Y_i - \alpha - \tau_h D_i - \beta(X_i - c) - \gamma D_i(X_i - c)\big)^2.$$

これはカーネル $k(u) = \frac{1}{2}\mathbb{1}\{|u| \le 1\}$ を用いた LL 推定量での **(3.16) 式** 推定と等価. 関数のクラスは以下:

$$\mathcal{M}(K) = \{\mu : |\mu'(a) - \mu'(b)| \le K|a - b| \text{ for all } a, b < c \text{ and } a, b > c\}. \tag{3.18}$$

ここで, K **はサンプリングパラメータ**. さらに次のような定義を導入する.

$$\tilde{\tau}_h = \mathbb{E}[\widehat{\tau}_h \mid X_1, \dots, X_{n_h}] \quad \text{where } n_h = \sum_{i=1}^{n} \mathbb{1}\{|X_i - c| \le h\}$$

$\mathrm{Var}[\widehat{\tau}_h \mid X_1, \dots, X_{n_h}]$ の nearest-neighbor 推定量を $\widehat{\sigma}_{NN}$ とすると, t 統計量は次のように分解できる:

$$\sqrt{n_h}\frac{\widehat{\tau}_h - \tau_{\mathrm{rdd}}}{\widehat{\sigma}_{NN}} = \sqrt{n_h}\frac{\widehat{\tau}_h - \tilde{\tau}_h}{\widehat{\sigma}_{NN}} + \sqrt{n_h}\frac{\tilde{\tau}_h - \tau_{\mathrm{rdd}}}{\widehat{\sigma}_{NN}}.$$

$$\sqrt{n_h}\frac{\widehat{\tau}_h - \tilde{\tau}_h}{\widehat{\sigma}_{NN}} \xrightarrow{d} \mathcal{N}(0,1), \quad \gamma_{\sup} \equiv \sup_{\mu \in \mathcal{M}(K)} \sqrt{n_h}\frac{|\tilde{\tau}_h - \tau_{\mathrm{rdd}}|}{\widehat{\sigma}_{NN}}.$$

$h \to 0$ **は要求されず**, $n_h \to \infty$ さえ成り立っていればよい. $|\mathcal{N}(\gamma, 1)|$ の $1 - \alpha$ 分位点を $cv_{1-\alpha}(\gamma)$ とすれば:

$$\inf_{\mu \in \mathcal{M}(K)} \mathbf{P}_\mu\left(\left|\sqrt{n_h}\frac{\widehat{\tau}_h - \tau_{\mathrm{rdd}}}{\widehat{\sigma}_{NN}}\right| \le cv_{1-\alpha}(\gamma_{\sup})\right) \ge 1 - \alpha$$

が成り立つ. したがって, 次の区間

$$\mathrm{CI} := \left[\widehat{\tau}_h - cv_{1-\alpha}(\gamma_{\sup})\frac{\widehat{\sigma}_{NN}}{\sqrt{n_h}},\ \widehat{\tau}_h + cv_{1-\alpha}(\gamma_{\sup})\frac{\widehat{\sigma}_{NN}}{\sqrt{n_h}}\right]$$

は honest な信頼区間となる.

まとめ: Kolesar and Rothe (2018)

- **メリット:**
  - バンド幅に関して $h \to 0$ の仮定が不要 (スコアが離散で適用可能)
  - 信頼区間を最も狭くするようにバンド幅を選べる
- **デメリット:**
  - K を研究者が自分で設定する必要がある
  - K をデータから推定した場合, 信頼区間は honest ではなくなる

澤田先生曰く, ノンパラの界隈からは評判悪い手法らしい. なので結果的には Calonico et al. (2014) に則って `rdrobust` で推定を行うのが一般的.

■visualization　末石 ミクロ計量分析 参照. 実際の分析とは異なるビジュアル化であることに注意せよ.

4　セミパラメトリック回帰

■p.73 **モチベーション**　Chernozhukov et al. (2018) でもセミパラの文献が多く引用されていることからも分かるように, ML の文脈とも深く関わりのある分野であるため, 特に漸近分布の導出 (**4.4 節**) は重要である.

■p.74 **セミパラの定義**　本書では以下のように定義して考察範囲を限定している:

Def: セミパラメトリックモデル

有限次元のパラメータの推定過程で, 未知関数をノンパラメトリックに推定しなければならないモデル

線形回帰モデルは誤差項の分布は推定しなくて良いため, この範疇ではない.

4.1 部分回帰モデル

■p.74 **定式化**　以下のように定式化される:

$$Y_i = \mathbf{X}_i^\top \boldsymbol{\beta} + g(Z_i) + e_i, \quad \mathbb{E}[e_i \mid \mathbf{X}_i, Z_i] = 0 \tag{4.1}$$

ノンパラパートでは次元の呪いを回避するため, Z_i の次元はせいぜい 3. 局外パラメータは Z_i に入れ込まれる. 線形パートには定数項が入っていないことに注意せよ.

■p.75 Robinson (1988)　$Y_i - \mathbb{E}[Y_i \mid Z_i = z]$ を $\mathbf{X}_i - \mathbb{E}[\mathbf{X}_i \mid Z_i = z]$ に回帰して得られる $\widehat{\boldsymbol{\beta}}_{\text{inf}}$ は, 条件付期待値を推定したものに置き換えた, $Y_i - \widehat{\mathbb{E}[Y_i \mid Z_i = z]}$ の $\mathbf{X}_i - \widehat{\mathbb{E}[\mathbf{X}_i \mid Z_i = z]}$ による回帰で得られる $\widehat{\boldsymbol{\beta}}$ と同じ漸近分布に分布収束する. また, $\widehat{\boldsymbol{\beta}}$ は $\sqrt{n}$ 一致性を満たす.

4.2 シングルインデックスモデル

■p.77 Ichimura (1993)　シングルインデックスモデルは以下で与えられる:

$$Y_i = g(\mathbf{X}_i^\top \boldsymbol{\beta}) + e_i, \qquad \mathbb{E}[e_i \mid \mathbf{X}_i] = 0.$$

ここで $g(\cdot)$ は **未知** の関数で, $\boldsymbol{\beta}$ は識別のため第一成分を 1 に基準化し, $\mathbf{X}_i^\top \boldsymbol{\beta} = X_{i1} + \tilde{\mathbf{X}}_i^\top \boldsymbol{\delta}$, $\boldsymbol{\beta} = [1, \boldsymbol{\delta}^\top]^\top$. 問題は, $g(\cdot)$ をノンパラメトリックに推定する際に, 未知パラメータ $\boldsymbol{\beta}$ に依存する条件付き期待値

$$g(\mathbf{x}^\top \boldsymbol{\beta}) = \mathbb{E}[Y_i \mid \mathbf{X}_i^\top \boldsymbol{\beta} = \mathbf{x}^\top \boldsymbol{\beta}]$$

を直接推定することの困難さにある. この回避のため, 任意のパラメータベクトル $\mathbf{b}$ に対して次を定義する:

$$G(\mathbf{x}^\top \mathbf{b}) \equiv \mathbb{E}[Y_i \mid \mathbf{X}_i^\top \mathbf{b} = \mathbf{x}^\top \mathbf{b}] \tag{4.5}$$

すなわち, 未知の $\boldsymbol{\beta}$ を $\mathbf{b}$ で **代用** する. $\mathbf{b}$ を与えれば, 標本 $\{Y_i, \mathbf{X}_i^\top \mathbf{b}\}_{i=1}^n$ から一個抜き NW 推定量

$$\widehat{G}_{-i}(\mathbf{x}^\top \mathbf{b}) = \frac{\sum_{j \neq i} k(\frac{\mathbf{X}_j^\top \mathbf{b} - \mathbf{x}^\top \mathbf{b}}{h}) Y_j}{\sum_{j \neq i} k(\frac{\mathbf{X}_j^\top \mathbf{b} - \mathbf{x}^\top \mathbf{b}}{h})}$$

で G を推定できる, このノンパラメトリック推定量を用いれば, $\boldsymbol{\delta}$ $(\boldsymbol{\beta})$ の推定が行える: [*11]

$$\widehat{\boldsymbol{\delta}} = \arg\min_{\mathbf{d}} \sum_{i=1}^n \big(Y_i - \widehat{G}_{-i}(X_{i1} + \tilde{\mathbf{X}}_i^\top \mathbf{d})\big)^2,$$

■p.78 **次元の呪いの回避**　$k(\frac{\mathbf{X}_j^\top \mathbf{b} - \mathbf{x}^\top \mathbf{b}}{h})$ の一次元のみを利用していることで, 次元の呪いを回避可能.

[*11] 確率的に十分な密度がある点のみを推定に用いる, トリミングを行うのがより正確.

■p.79 $\boldsymbol{\delta}$ **の漸近分布**　適当な条件の下で, 漸近分布は次のようになる:

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{d} N\big(\mathbf{0}, \mathbf{Q}^{-1}\boldsymbol{\Omega}\mathbf{Q}^{-1}\big) \quad \text{where } \mathbf{Q}=\mathbb{E}\big[\tilde{\mathbf{X}}_i^*\tilde{\mathbf{X}}_i^{*\top}\big],\ \boldsymbol{\Omega}=\mathbb{E}\big[\tilde{\mathbf{X}}_i^*\tilde{\mathbf{X}}_i^{*\top}e_i^2\big],$$

■**参考**: **デルタ法**　証明に **デルタ法** を使うのかと思ったら違うみたい. デルタ法は以下:

定理: Delta Method (デルタ法)

線形回帰モデル $Y=\mathbf{X}^\top\boldsymbol{\beta}+\mathbf{e}$ を考え, 適切な仮定の結果, 漸近的なふるまいとして

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, \mathbf{Q}^{-1}\boldsymbol{\Omega}\mathbf{Q}^{-1}) \quad \text{where } \mathbf{Q}=\mathbb{E}[\mathbf{X}_i\mathbf{X}_i^\top],\ \boldsymbol{\Omega}=\mathbb{E}[\mathbf{X}_i\mathbf{X}_i^\top e_i^2],$$

が成立する. ここで $\mathbf{f}(\widehat{\boldsymbol{\beta}})$ を $\widehat{\boldsymbol{\beta}}$ の C^1 級関数の集合とすると, 漸近分布と漸近共分散行列推定量は:

$$\sqrt{n}\big(\mathbf{f}(\widehat{\boldsymbol{\beta}})-\mathbf{f}(\boldsymbol{\beta})\big) \xrightarrow{d} \mathcal{N}\left(\mathbf{0},\ \left(\frac{\partial f(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^\top}\right)\mathbf{Q}^{-1}\boldsymbol{\Omega}\mathbf{Q}^{-1}\left(\frac{\partial f(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^\top}\right)^\top\right).$$

$$\widehat{\mathrm{AVar}}\big(\mathbf{f}(\widehat{\boldsymbol{\beta}})\big) = \left(\frac{\partial f(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^\top}\mid_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\right)\widehat{\mathbf{Q}^{-1}\boldsymbol{\Omega}\mathbf{Q}^{-1}}\left(\frac{\partial f(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^\top}\mid_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\right)^\top.$$

Proof 上級計量 **Ch.4** に則って証明する. 平均値の定理により

$$f(\widehat{\boldsymbol{\beta}}) = f(\boldsymbol{\beta}) + \frac{\partial f(\bar{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}^\top}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}),$$

ただし $\bar{\boldsymbol{\beta}}$ は $\widehat{\boldsymbol{\beta}}$ と $\boldsymbol{\beta}$ の間にある点. $\widehat{\boldsymbol{\beta}}\xrightarrow{p}\boldsymbol{\beta}$ から $\bar{\boldsymbol{\beta}}\xrightarrow{p}\boldsymbol{\beta}$ であるため, 連続写像定理により

$$\sqrt{n}\big(\mathbf{f}(\widehat{\boldsymbol{\beta}})-\mathbf{f}(\boldsymbol{\beta})\big) \xrightarrow{d} \mathcal{N}\left(\mathbf{0},\ \left(\frac{\partial f(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^\top}\right)\mathbf{Q}^{-1}\boldsymbol{\Omega}\mathbf{Q}^{-1}\left(\frac{\partial f(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^\top}\right)^\top\right).$$

漸近共分散行列推定量は置き換えと代入により得られる. □

■p.79 **漸近分布の導出**　では実際に導出をしてみよう. アイデア自体はデルタ法と同じだ:

Proof 目的関数を

$$S_n(\mathbf{d}) \ = \ \sum_{i=1}^n\big(Y_i-\widehat{G}_{-i}(X_{i1}+\tilde{\mathbf{X}}_i^\top\mathbf{d})\big)^2 \ = \ \sum_{i=1}^n\big(Y_i-\widehat{G}_{-i}(t_i(\mathbf{d}))\big)^2$$

また重要な量を以下のように記述する:

$$\tilde{\mathbf{X}}_i^* \coloneqq \frac{\partial}{\partial\mathbf{d}}G\big(X_{i1}+\tilde{\mathbf{X}}_i^\top\mathbf{d}\big)\mid_{\mathbf{d}=\boldsymbol{\delta}} \coloneqq G_i'\,\tilde{\mathbf{X}}_i,\ \mathbf{Q}=\mathbb{E}\big[\tilde{\mathbf{X}}_i^*\tilde{\mathbf{X}}_i^{*\top}\big],\ \boldsymbol{\Omega}=\mathbb{E}\big[\tilde{\mathbf{X}}_i^*\tilde{\mathbf{X}}_i^{*\top}e_i^2\big]$$

仮定は以下の通り:

(A1) $\widehat{G}_{-i}$, G は十分に滑らか,　(A2) $h\to0$, $nh\to\infty$ 等のバンド幅条件
(A3) leave-one-out を行い, 密度は有界,　(A4) 適当なモーメント条件 (4 次モーメント等)

FOC は連鎖律より

$$\frac{\partial S_n(\mathbf{d})}{\partial \mathbf{d}} = -2\sum_{i=1}^{n}\big(Y_i - \widehat{G}_{-i}(t_i(\mathbf{d}))\big)\,\frac{\partial \widehat{G}_{-i}(t_i(\mathbf{d}))}{\partial \mathbf{d}}$$
$$= -2\sum_{i=1}^{n}\big(Y_i - \widehat{G}_{-i}(t_i(\mathbf{d}))\big)\,\widehat{G}'_{-i}(t_i(\mathbf{d}))\,\tilde{\mathbf{X}}_i = \mathbf{0}$$

$\widehat{\boldsymbol{\delta}}$ は FOC を満たすため $\dfrac{\partial S_n(\widehat{\boldsymbol{\delta}})}{\partial \mathbf{d}} = \mathbf{0}$ が成り立ち，これを真値 $\boldsymbol{\delta}$ の近傍でテイラー展開すると:

$$\mathbf{0} = \frac{\partial S_n(\boldsymbol{\delta})}{\partial \mathbf{d}} + \frac{\partial^2 S_n(\bar{\boldsymbol{\delta}})}{\partial \mathbf{d}\partial \mathbf{d}^\top}(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}),$$

$\bar{\boldsymbol{\delta}}$ は $\widehat{\boldsymbol{\delta}}$ と $\boldsymbol{\delta}$ の間の点．$\mathbf{d} = \boldsymbol{\delta}$ を代入し，$Y_i - \widehat{G}_{-i}(t_i(\boldsymbol{\delta}))$ を次のように分解する:

$$Y_i - \widehat{G}_{-i}(t_i(\boldsymbol{\delta})) = \underbrace{Y_i - G(t_i(\boldsymbol{\delta}))}_{=e_i} + \underbrace{G(t_i(\boldsymbol{\delta})) - \widehat{G}_{-i}(t_i(\boldsymbol{\delta}))}_{\text{バイアス}}.$$

$$\frac{1}{2}\frac{\partial S_n(\boldsymbol{\delta})}{\partial \mathbf{d}} = -\sum_{i=1}^{n} e_i\,\widehat{G}'_{-i}(t_i(\boldsymbol{\delta}))\,\tilde{\mathbf{X}}_i - \sum_{i=1}^{n}\big(G(t_i(\boldsymbol{\delta})) - \widehat{G}_{-i}(t_i(\boldsymbol{\delta}))\big)\,\widehat{G}'_{-i}(t_i(\boldsymbol{\delta}))\,\tilde{\mathbf{X}}_i.$$

条件により次が成り立つ，上の評価が甘そうだな:

$$\widehat{G}'_{-i}(t_i(\boldsymbol{\delta})) = G'(t_i(\boldsymbol{\delta})) + o_p(1)$$

$$\sum_{i=1}^{n}\big(G(t_i(\boldsymbol{\delta})) - \widehat{G}_{-i}(t_i(\boldsymbol{\delta}))\big)\,\widehat{G}'_{-i}(t_i(\boldsymbol{\delta}))\,\tilde{\mathbf{X}}_i = o_p(\sqrt{n}).$$

したがって，分解した式に代入すると:

$$\frac{1}{\sqrt{n}}\frac{\partial S_n(\boldsymbol{\delta})}{\partial \mathbf{d}} = -2\frac{1}{\sqrt{n}}\sum_{i=1}^{n} G'(t_i(\boldsymbol{\delta}))\,\tilde{\mathbf{X}}_i\,e_i + o_p(1) = \frac{-2}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\mathbf{X}}_i^*\,e_i + o_p(1).$$

テイラー展開の 2 階導関数を展開すると主要項は

$$\frac{\partial^2 S_n(\mathbf{d})}{\partial \mathbf{d}\partial \mathbf{d}^\top} = 2\sum_{i=1}^{n}\big(\widehat{G}'_{-i}(t_i(\mathbf{d}))\big)^2\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i^\top \;+\; R_n(\mathbf{d}),$$

ここで $R_n(\mathbf{d})$ は残りの項（残差や $\widehat{G}''$ を含む項）で，仮定の下に $R_n(\bar{\boldsymbol{\delta}})/n = o_p(1)$ より:

$$\frac{1}{2n}\frac{\partial^2 S_n(\bar{\boldsymbol{\delta}})}{\partial \mathbf{d}\partial \mathbf{d}^\top} = \frac{1}{n}\sum_{i=1}^{n}\big(\widehat{G}'_{-i}(t_i(\boldsymbol{\delta}))\big)^2\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i^\top + o_p(1) \xrightarrow{p} \mathbb{E}\big[G'(t_i(\boldsymbol{\delta}))^2\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i^\top\big] = \mathbb{E}\big[\tilde{\mathbf{X}}_i^*\tilde{\mathbf{X}}_i^{*\top}\big] = \mathbf{Q}.$$

以上の結果を使いテイラー展開を解くと

$$\sqrt{n}(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}) = -\mathbf{Q}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\mathbf{X}}_i^* e_i + o_p(1).$$

i.i.d. 性とモーメント条件の下で，中心極限定理により:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\mathbf{X}}_i^* e_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}) \quad \boldsymbol{\Omega} = \mathbb{E}\big[\tilde{\mathbf{X}}_i^*\tilde{\mathbf{X}}_i^{*\top}e_i^2\big].$$

$$\sqrt{n}(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}) \xrightarrow{d} N\big(\mathbf{0}, \mathbf{Q}^{-1}\boldsymbol{\Omega}\mathbf{Q}^{-1}\big) \quad \text{where } \mathbf{Q} = \mathbb{E}\big[\tilde{\mathbf{X}}_i^*\tilde{\mathbf{X}}_i^{*\top}\big],\ \boldsymbol{\Omega} = \mathbb{E}\big[\tilde{\mathbf{X}}_i^*\tilde{\mathbf{X}}_i^{*\top}e_i^2\big],$$

を得る．これは目標の関係式である．　□

松下先生曰く, 掛け算のオーダーを考える時, 片方が 0 に確率収束するような時には, もう片方のオーダーは十分に大きくなるので, ある程度ラフに仮定を置いて抑えてしまっても問題ないらしい.

■デルタ法との差異 デルタ法は，もしさらに $\boldsymbol{\delta}$ が別のパラメータ $\boldsymbol{\beta}$ の滑らかな関数で定義される場合（$\boldsymbol{\delta} = f(\boldsymbol{\beta})$）に, 漸近分散を求めるため用いるもので，**今回の手法とは異なる**.

■p.79 Ichimura (1993) の問題点 $\widehat{\boldsymbol{\delta}}$ の漸近分布は, G が既知であるときの NL(L)S 推定量の漸近分布と等しい. しかし, 本来の推定対象であった g が既知であるときの NLS 推定量の漸近分布とは漸近分散が異なる. おそらく効率性が下がるのだろう.

4.3 平均処置効果

■p.79 ATE の傾向スコアを利用した識別 1.4 節 のとおり, ATE は強い無視可能性の下で識別でき:

$$\tau_{\text{ate}} = \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbf{E}[\mu_1(\mathbf{X}_i) - \mu_0(\mathbf{X}_i)] \quad \text{where } \mu_d(\mathbf{x}) \coloneqq \mathbb{E}[Y_i \mid \mathbf{X}_i = \mathbf{x}, D_i = d] \tag{4.6}$$

と書き換えられた. さらに **同様の仮定の下で** 以下が成り立つ:

$$\begin{aligned}\eta_d(\mathbf{x}) &\coloneqq \mathbb{E}[D_i Y_i \mid \mathbf{X}_i = \mathbf{x}] = \mathbb{E}[D_i Y_i(1) \mid \mathbf{X}_i = \mathbf{x}] = p(\mathbf{x})\mu_1(\mathbf{x}) \\ \eta_{1-d}(\mathbf{x}) &\coloneqq \mathbb{E}[(1-D_i) Y_i \mid \mathbf{X}_i = \mathbf{x}] = \mathbb{E}[(1-D_i) Y_i(0) \mid \mathbf{X}_i = \mathbf{x}] = (1-p(\mathbf{x}))\mu_0(\mathbf{x})\end{aligned}$$

ここで $p(\mathbf{x}) = \mathbf{P}(D_i = 1 \mid \mathbf{X}_i = \mathbf{x})$ は傾向スコア.

Proof ここでは $\eta_d(\mathbf{x})$ についてのみ示す. まず LIE より:

$$\eta_d(\mathbf{x}) = \mathbb{E}[D_i Y_i(1) \mid \mathbf{X}_i = \mathbf{x}] = \mathbb{E}[D_i \, \mathbb{E}[Y_i(1) \mid D_i, \, \mathbf{X}_i = \mathbf{x}] \mid \mathbf{X}_i = \mathbf{x}] \, .$$

強い無視可能性 (非交絡の仮定) より $\mathbb{E}[Y_i(1) \mid D_i, \mathbf{X}_i = \mathbf{x}] = \mathbb{E}[Y_i(1) \mid \mathbf{X}_i = \mathbf{x}] = \mu_1(\mathbf{x})$ が成り立ち:

$$\eta_d(\mathbf{x}) = \mathbb{E}[D_i \mu_1(\mathbf{x}) \mid \mathbf{X}_i = \mathbf{x}] = \mu_1(\mathbf{x}) \, \mathbb{E}[D_i \mid \mathbf{X}_i = \mathbf{x}] = p(\mathbf{x}) \, \mu_1(\mathbf{x}).$$

以上より題意の式を得た. □

上記の変形を利用して以下が得られる:

$$\tau_{\text{ate}} = \mathbb{E}\left[\frac{\eta_d(\mathbf{X}_i)}{p(\mathbf{X}_i)} - \frac{\eta_{1-d}(\mathbf{X}_i)}{1-p(\mathbf{X}_i)}\right] \tag{4.7}$$

さらに, 非交絡の仮定から以下が示せる:

$$\mathbb{E}\left[\frac{D_i Y_i}{p(\mathbf{X}_i)}\right] = \mathbb{E}[Y_i(1)], \quad \mathbb{E}\left[\frac{(1-D_i) Y_i}{1-p(\mathbf{X}_i)}\right] = \mathbb{E}[Y_i(0)].$$

Proof まず $D_i Y_i = D_i Y_i(1)$ より

$$\mathbb{E}\left[\frac{D_i Y_i}{p(\mathbf{X}_i)}\right] = \mathbb{E}\left[\frac{D_i Y_i(1)}{p(\mathbf{X}_i)}\right] \overset{\text{LIE}}{=} \mathbb{E}\left[\mathbb{E}\left[\frac{D_i Y_i(1)}{p(\mathbf{X}_i)} \,\middle|\, \mathbf{X}_i\right]\right],$$

$$\mathbb{E}\left[\frac{D_i Y_i(1)}{p(\mathbf{X}_i)} \mid \mathbf{X}_i\right] = \frac{\mathbb{E}[D_i \mid \mathbf{X}_i] \, \mathbb{E}[Y_i(1) \mid \mathbf{X}_i]}{p(\mathbf{X}_i)} = \mathbb{E}[Y_i(1) \mid \mathbf{X}_i] \quad (\because \ (Y_i\,(1), \, Y_i\,(0)) \perp\!\!\!\perp D_i \mid \mathbf{X}_i),$$

したがって:

$$\mathbb{E}\left[\frac{D_i Y_i}{p(\mathbf{X}_i)}\right] = \mathbb{E}[\mathbb{E}[Y_i(1) \mid \mathbf{X}_i]] = \mathbb{E}[Y_i(1)].$$

を得る, $Y_i(0)$ についても同様. □

この結果を用いれば, ATE を以下のように識別することも可能である:

$$\tau_{\text{ate}} = \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}\left[\frac{D_i Y_i}{p(\mathbf{X}_i)} - \frac{(1-D_i)Y_i}{1-p(\mathbf{X}_i)}\right] \tag{4.8}$$

■p.80 **傾向スコア定理** 傾向スコアには以下のような性質がある:

Theorem: 傾向スコア定理

非交絡の仮定が成立しているとき, 以下が成り立つ:

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp D_i \mid p(\mathbf{X}_i)$$

Proof 任意の値 y_1, y_0 と $d \in \{0,1\}$ をとる。まず，条件付き確率を X_i を介して分解する：

$$\begin{aligned}&\mathbf{P}(Y_i(1) = y_1,\, Y_i(0) = y_0 \mid D_i = d,\, e(X_i) = e)\\ =&\mathbb{E}[\mathbf{P}(Y_i(1) = y_1, Y_i(0) = y_0 \mid D_i = d, X_i) \mid D_i = d,\, e(X_i) = e]\,.\end{aligned}$$

非交絡より内側の確率は D_i に依存しない：

$$\mathbf{P}(Y_i(1) = y_1, Y_i(0) = y_0 \mid D_i = d, X_i) = \mathbf{P}(Y_i(1) = y_1, Y_i(0) = y_0 \mid X_i).$$

したがって

$$\mathbf{P}(Y_i(1) = y_1, Y_i(0) = y_0 \mid D_i = d, e(X_i) = e) = \mathbb{E}[\mathbf{P}(Y_i(1) = y_1, Y_i(0) = y_0 \mid X_i) \mid D_i = d,\, e(X_i) = e]\,.$$

ここで $e(X_i)$ は X_i の関数であり，条件 $e(X_i) = e$ のもとでは X_i の集合は「$e(X_i) = e$ を満たす X_i」に制限される。この制約は $D_i = d$ の条件と無関係であるため，

$$\begin{aligned}&\mathbb{E}[\mathbf{P}(Y_i(1) = y_1, Y_i(0) = y_0 \mid X_i) \mid D_i = d,\, e(X_i) = e]\\ =&\mathbb{E}[\mathbf{P}(Y_i(1) = y_1, Y_i(0) = y_0 \mid X_i) \mid e(X_i) = e]\,.\end{aligned}$$

右辺はまさに

$$\mathbf{P}(Y_i(1) = y_1, Y_i(0) = y_0 \mid e(X_i) = e).$$

以上より

$$\mathbf{P}(Y_i(1) = y_1, Y_i(0) = y_0 \mid D_i = d, e(X_i) = e) = \mathbf{P}(Y_i(1) = y_1, Y_i(0) = y_0 \mid e(X_i) = e),$$

したがって $(Y_i(1), Y_i(0)) \perp D_i \mid e(X_i)$ が示された。 □

非交絡の仮定は共変量の値が同じもの同士での比較が可能なことを意味するが, 傾向スコア定理は, 傾向スコアの値が同じもの同士での比較が可能なことを意味する. 図的なイメージは **林 (2024)** から拝借しよう:

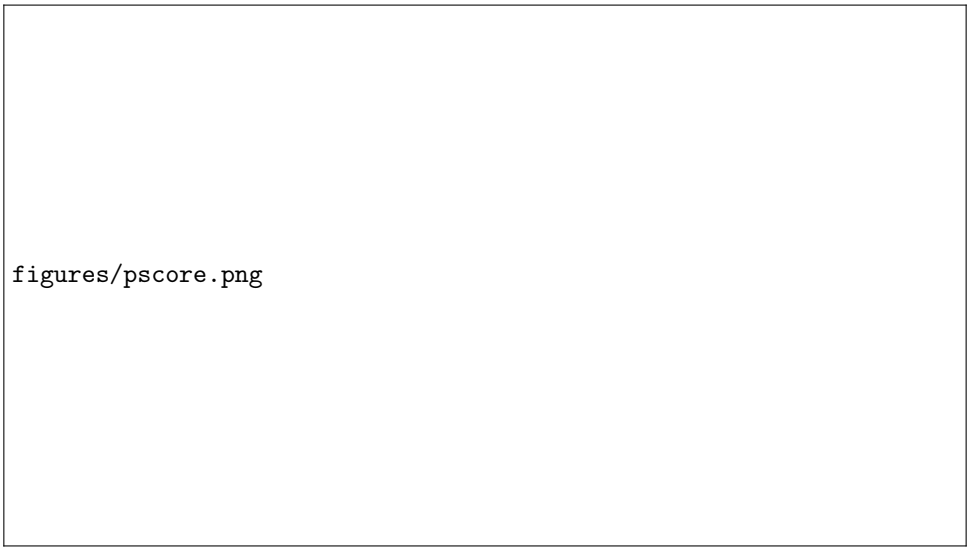

図 4　傾向スコア定理の意味

4.4　プラグイン推定量の漸近分布*

4.5　補論**

5　決定木とアンサンブル学習

5.1　決定木

5.1.1　分類木

■p.99 **ジニ係数**　Gini's Diversity Index または ジニ不純度, Gini Coefficient とは違うものらしい. 最大値が 1 でないことからも明らかだろう. CART は各ノードにおいて不純度 (の加重平均) を最小にする (うまく分けられる) 分岐を作る (共変量と閾値を選択する) ので, 局所最適ではあるが全体最適である保証はない.

■p.103 **モデルの複雑さ**　共変量は再利用できるが, 本来の目的は予測にあるためモデルの複雑さとはトレードオフがあるべき.

5.1.2　回帰木

■p.104 **損失関数**　各ノードの標本平均の加重和 (に分割前の観測個数を掛けたもの) を損失関数とする.

■p.104 **回帰としての解釈**　回帰関数の推定量は階段関数となる. 分岐を増やすことはカーネル推定におけるバンド幅を小さくすることに対応するため, ノンパラと同様の理屈で分散が大きくなる.

■p.105 **剪定 (prune)**　リーフを十分に細分化する回帰木 T_0 から, 不要なエッジ (分岐) を剪定してリーフを減らす. 基準として以下を用いる:

$$C_\alpha(T) = \sum_{i=1}^{n} (Y_i - \widehat{\mu}_T(\mathbf{X}_i))^2 + \alpha|T|$$

ここで, $|T|$ はリーフの数, $\alpha > 0$ はチューニングパラメータ. α は交差検証法などで決定される.

■なぜ交差検証法か? T_0 から得られる部分木は有限個であり，複雑度基準は各木 T で α に関して線形ゆえ

$$0 = \alpha_0 < \alpha_1 < \cdots < \alpha_K < \infty$$

が存在して,

$$\alpha \in [\alpha_k, \alpha_{k+1}) \quad \Rightarrow \quad \arg\min_T C_\alpha(T) = T_k$$

となるような有限個の部分木 $T_0 \supset T_1 \supset \cdots \supset T_K$ が得られる. このため，α を連続的に選択する必要はなく，この包含列に属する有限個の部分木を候補として選択すれば十分である. これらの部分木について k-fold の交差検証法をおこなうことによって最適な T_k, そしてそれに対応する α (の範囲) を決定することが出来る.

■p.105 次元の呪いの回避　回帰木は，予測に寄与しない共変量をデータ依存的に用いないという選択を通じて，暗黙の次元削減を行うため，共変量の次元がある程度大きくても予測が可能である。一方で，不要な共変量に対してバンド幅を十分に大きく設定できるならば，カーネル推定量も次元削減と同様の振る舞いを示すように思われる。しかし，カーネル推定におけるバンド幅 h は MISE 最小化という目的関数のもとで，評価点 $\mathbf{x}$ に依らない共通の平滑化パラメータとして選択される。その結果，$h \to \infty$ のような極端な値は全体のバイアスを増大させるため最適化の候補から排除され，回帰木のように局所的・次元ごとに平滑化を切り替えることはできない。

■p.105 CART のバイアス (Athey and Imbens, 2016)　簡潔に言えば, 分布にカットオフ c を入れることとなって母集団平均が変わるためにバイアスが生じる. 回避のためには, 分割と推定に用いる観測値を分けるべき (honest estimation). これならば分割がカットオフとして機能しない.

■p.106 代替的なリーフの構築法 (Athey and Imbens, 2016)　Honest estimation のためには CART は不適切. Athey and Imbens (2016) は代替的な手法を提案し, また因果推論の問題に拡張して CATE を推定する方法も提案している. これは処置効果の異質性を捉えるために有用である (Ch 5.4).

5.2 バギングとランダムフォレスト

■p.107 アンサンブル学習　単体では性能の良くないアルゴリズムを組み合わせてよい予測を得るテクニック.

■p.107 バギング (bootstrap aggregating, Breiman, 1996)　応答変数と共変量のペアを復元抽出した, ブートストラップ標本を用いて推定値の構成を繰り返し, その平均を最終的な推定値とする手法. 分散を削減できるが, その効果は回帰木同士の相関が高い (複製サンプルを用いているために似通った形になる) ことで限定的になるのが弱点といえる.

■p.108 ランダムフォレスト (Breiman, 2001)　バギングの手順のうち, ブートストラップ[*12] 標本を用いた推定値の構成において, 利用する共変量もランダムに選択する手法. バギングの一般化といえる. 加重平均としての解釈は Σ の交換に注意すれば確認できる. また, 回帰関数の書き方には一定の注意が必要だろう:

$$\widehat{\mu}(\mathbf{x}) = \sum_{j=1}^{J} \frac{\sum_{i=1}^{n} \mathbb{1}\{\mathbf{X}_i \in L_j\} Y_i}{\sum_{i=1}^{n} \mathbb{1}\{\mathbf{X}_i \in L_j\}} \mathbb{1}\{\mathbf{x} \in L_j\} \tag{5.2}$$

[*12] サブサンプリングでも良い.

$$\widehat{\mu}_b(\mathbf{x}) = \frac{\sum_{i=1}^n \mathbb{1}\{\mathbf{X}_i \in L_b(\mathbf{x})\} Y_i}{\sum_{i=1}^n \mathbb{1}\{\mathbf{X}_i \in L_b(\mathbf{x}\}}$$

この二式は同じ意味を持っている. ここで $\mathbf{x}$ を含むリーフを $L_b(\mathbf{x})$ としている. その複雑さから, あまり数学的な性質の解析は進んでいないようだ.

Proof 回帰木による推定量は，リーフ集合 $\{L_j\}_{j=1}^J$ を用いて

$$\hat{\mu}(\mathbf{x}) = \sum_{j=1}^{J} \left(\frac{\sum_{i=1}^n \mathbb{1}\{\mathbf{X}_i \in L_j\} Y_i}{\sum_{i=1}^n \mathbb{1}\{\mathbf{X}_i \in L_j\}} \right) \mathbb{1}\{\mathbf{x} \in L_j\}$$

と書くことができる。ここで $\{L_j\}$ は互いに素な分割であるため，任意の $\mathbf{x}$ に対して $\mathbb{1}\{\mathbf{x} \in L_j\}$ はただ一つの $j = j(\mathbf{x})$ でのみ 1 を取る。したがって，上式は

$$\hat{\mu}(\mathbf{x}) = \frac{\sum_{i=1}^n \mathbb{1}\{\mathbf{X}_i \in L_{j(\mathbf{x})}\} Y_i}{\sum_{i=1}^n \mathbb{1}\{\mathbf{X}_i \in L_{j(\mathbf{x})}\}}$$

に簡約される。ここで $L_{j(\mathbf{x})}$ は $\mathbf{x}$ を含む唯一のリーフであり，これを $L_b(\mathbf{x})$ と書けば，両者は同一の推定量を異なる記法で表したものに過ぎない。 □

5.3 ブースティング

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.

■独立性　⫫ の記号で表す.

■事象　大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.

■確率変数の値域, 集合族, 事象族　カリグラフィー `mathcal` 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, $\mathbf{W}$ の標本空間の元についてこの関係が成立する場合, 単に $\mathbb{E}[Y \mid S, \mathbf{W}]$ 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.

■対角行列　$\mathrm{diag}(a_1, \ldots, a_k)$ と対角成分を具体的に記載して表現する.

■定義　$:=$ を用いて記載する.

付録 B　数学的準備 (書評: 『計量経済学のための数学』)

B.1　『計量経済学のための数学』

■書評的なもの　本書では数学的補遺が特に準備されていない. 数学的な準備のため, 『計量経済学のための数学』を用いて数学的準備を行うことにした. この本の前書きにはこのような記載がある:

田中 久稔, 計量経済学のための数学, p.i

> 『本書では, 計量経済学の基礎を理解するために必要な最低限のトピックを厳選し, それらについてのみ集中的に解説することを心がけました. 標準的な線形代数や確率論のテキストであれば必ず扱うべき内容 (「クラメールの公式」, **「基本変形」**, 「ジョルダン標準形」, etc.) であっても, 計量経済学とのかかわりが薄ければ, ばっさりとカットしています.』 (強調は前川)

実際宣言通りの内容である. 計量に使う数学として必要な内容を最短経路で提示し, かつ無駄が極限までそぎ落とされている, 中々に素晴らしい内容の本だ. 1 週間もあれば読破可能なのでぜひ手に取ってみてほしい.

■線形写像としての行列　上記の内容から分かるように, 本書は基本変形 (掃き出し法, ガウスの消去法) から出発する一次連立方程式システムを中心とした構成ではない. この点で, 自身が把握している線形代数を扱う他の教科書 (三宅, 齋藤, Simon & Blume) とは大きな違いがある. 一貫して以下の観点で議論がすすむ:

田中 久稔, 計量経済学のための数学, p.43

> **『行列とは線形写像です.』**

行列の積は合成写像として定義されており, 逆行列の存在は逆写像の存在と同値な全単射と結びつけられる. ランクは像の次元として定義されており, 像とスパンの一致を根拠にして議論が進んでいく.

■射影行列, OLS の幾何的解釈　5 章では, 射影行列によってベクトルが, どのような部分空間に写されるのか (特に射影されるのか) を丁寧に説明してくれる. 図形的には, **図 4** のようにまとめられる. 他にも例えば, 定理 5.6 (2) は, 2SLS での第一段階回帰を経ても外生変数が保存されることの根拠となっており重要である.

図 5　直交分解のイメージ

■行列の平方根 (p.116, **問題** 6.6)　行列の平方根は, 存在性については, 練習問題 (すべて解答付き!) で確認することができる. 末石本では更に半正定値行列である平方根行列の一意性が (証明無しで) 言及されている.

Def: 平方根行列

$$\mathbf{B}^\top\mathbf{B} \coloneqq \mathbf{B}^2 = \mathbf{A}$$

となるような $\mathbf{B}$ を $\mathbf{A}$ の平方根と呼び, $\mathbf{A}^{1/2}$ と表記する.

Thm: 半正定値行列の平方根行列

$\mathbf{L}^\top\mathbf{L} = \mathbf{I}_k$ を満たす $\mathbf{L}$ と, 非負の相異なる実数対角成分を持つ対角行列 $\mathbf{\Lambda}$ によって, $\mathbf{A} = \mathbf{L}\mathbf{\Lambda}\mathbf{L}^\top$ と表現できる k 次対称行列 $\mathbf{A}$ については, 平方根行列 $\mathbf{A}^{1/2}$ が常に存在する.

Proof 『計量経済学のための数学』の定理 6.4 で存在が正当化される, 以下の行列の対角化を考える:

$$\mathbf{A} \;=\; \mathbf{L}\,\mathrm{diag}(\lambda_1,\ldots,\lambda_k)\,\mathbf{L}^\top$$

$\mathbf{L} \in \mathbb{R}^{k\times k}$ (対角化に用いる, 固有ベクトルを結合した行列) は直交行列で, $\mathbf{L}^\top\mathbf{L} = \mathbf{I}_k$ を満たす. $\mathrm{diag}(\lambda_1,\ldots,\lambda_k)$ の対角成分は $\lambda_i \geq 0$ を満たす. 系 6.2 から $\mathbf{A}$ は半正定値. $\mathrm{diag}(\lambda_1,\ldots,\lambda_k)$ に対し

$$\mathrm{diag}(\lambda_1,\ldots,\lambda_k)^{1/2} \coloneqq \mathrm{diag}(\sqrt{\lambda_1},\ldots,\sqrt{\lambda_k})$$

と定義すると, これは平方根行列であることを確認できる. $\mathrm{diag}(\lambda_1,\ldots,\lambda_k)^{1/2}$ も対角成分が非負であるから, 半正定値対称行列である. そこで, 対角行列の対称性と $\mathbf{L}^\top\mathbf{L} = \mathbf{I}_k$ を利用して:

$$\begin{aligned}\mathbf{A} \;&=\; \mathbf{L}\,\mathrm{diag}(\lambda_1,\ldots,\lambda_k)^{1/2}\,\mathrm{diag}(\lambda_1,\ldots,\lambda_k)^{1/2}\,\mathbf{L}^\top \\ &=\; \mathbf{L}\,\mathrm{diag}(\lambda_1,\ldots,\lambda_k)^{1/2}\,\mathbf{L}^\top\,\mathbf{L}\,\mathrm{diag}(\lambda_1,\ldots,\lambda_k)^{1/2}\,\mathbf{L}^\top \\ &=\; \left[\mathbf{L}\,\mathrm{diag}(\lambda_1,\ldots,\lambda_k)^{1/2}\,\mathbf{L}^\top\right]\left[\mathbf{L}\,\mathrm{diag}(\lambda_1,\ldots,\lambda_k)^{1/2}\,\mathbf{L}^\top\right]\end{aligned}$$

ここで, $\mathbf{A}^{1/2}$ を以下のように定義する:

$$\mathbf{A}^{1/2} \coloneqq \mathbf{L}\,\mathrm{diag}(\lambda_1,\ldots,\lambda_k)^{1/2}\,\mathbf{L}^\top\,.$$

すると, $\mathbf{A}^{1/2}$ は対称半正定値である. 対称行列であるから, 先の式変形により,

$$(\mathbf{A}^{1/2})^2 = \left[\mathbf{L}\,\mathrm{diag}(\lambda_1,\ldots,\lambda_k)^{1/2}\,\mathbf{L}^\top\right]\left[\mathbf{L}\,\mathrm{diag}(\lambda_1,\ldots,\lambda_k)^{1/2}\,\mathbf{L}^\top\right] = \mathbf{A}$$

が常に存在することが示された. 更にこの平方根行列は対称半正定値行列である. □

■回帰分析　以下の記述は, 必ず理解しておくべきだろう.

田中 久稔, 計量経済学のための数学, p.173-174

『回帰分析はよく知られた統計的方法の一つではありますが, 改めて「回帰分析とは何か, 30 字以内で説明せよ」と問われれば, 答えに窮する読者も多いのではないでしょうか? (中略) 「回帰分析とは何か」と問われて「最小二乗法のことである」と答えることも目的と手段を混同した誤りであるといえます. 』

田中 久稔, 計量経済学のための数学, p.174

『回帰分析とは, 「ある変数 X の値を手掛かりとして, 他の変数 Y の値を推測すること」』

■回帰係数の識別　まず識別の定義を確認する.

Def: 識別

興味の対象である θ を, 観測可能な量によって表現すること.

損失関数として $L(u) = u^2$ を採用し, かつ $\mathbb{E}[Y \mid \mathbf{X}] = \mathbf{X}^\top\boldsymbol{\beta}$ が成り立つ線形回帰モデルにおいて, 興味の対象である回帰係数 $\boldsymbol{\beta}$ は以下のように識別される:

Them 9.10: 回帰係数の識別

線形回帰モデルにおいて, $\mathbb{E}[\mathbf{X}\mathbf{X}^\top] \in \mathbb{R}^{k\times k}$ が正則であるならば, 以下が成り立つ:

$$\boldsymbol{\beta} = [\mathbb{E}[\mathbf{X}\mathbf{X}^\top]]^{-1}\mathbb{E}[\mathbf{X}Y]$$

■OLSE の構成　興味のある対象 θ を推定するために, 推測統計学では, 母集団上の確率変数を用いて推定量 $S_n = S(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ を構成する. 線形回帰モデルにおける $\boldsymbol{\beta}$ の推定量である, OLS 推定量 (OLSE) の構成の方法が本書では 2 通り示されている. 一つ目の方法は, 識別した回帰係数の期待値を標本対応させるというもの. 末石本では以下で説明する二つ目の方法を利用して OLSE を定義する. 特にこの二つ目の方法は, **MSE の記法に文献によって揺れがある理由**を与えてくれるため, 理解のため議論を整理しておこう.

Thm 9.11: 回帰係数の最良線形予測

線形回帰モデルで $\mathbb{E}[\mathbf{X}\mathbf{X}^\top]$ が正則ならば, $\mathrm{MSE}(\mathbf{b})$ を最小にする $\mathbf{b}$ は真の回帰係数 $\boldsymbol{\beta}$ に一致する, 即ち:

$$\arg\min_{\mathbf{b}} \mathbb{E}\big[(Y - \mathbf{X}^\top\mathbf{b})^2\big] = \boldsymbol{\beta}.$$

この結果を手掛かりに, LLN を利用して $\boldsymbol{\beta}$ を推定する方法が以下である.

Def: OLSE の間接的な構成

$\mathrm{MSE}(\mathbf{b}) = \mathbb{E}\big[(Y - \mathbf{X}^\top\mathbf{b})^2\big]$ をこの標本平均により**近似**したものを最小化するような $\mathbf{b}$:

$$\arg\min_{\mathbf{b}} \widehat{M}_n(\mathbf{b}) = \arg\min_{\mathbf{b}} \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i^\top\mathbf{b})^2, \text{ where } \widehat{M}_n(\mathbf{b}) \coloneqq \sum_{i=1}^{n}(Y_i - \mathbf{X}_i^\top\mathbf{b})^2$$

を OLSE $\widehat{\boldsymbol{\beta}}_n$ として定義する. $\mathbf{Y} \coloneqq (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$, $\mathbf{X} \coloneqq (\mathbf{X}_1, \ldots, \mathbf{X}_n)^\top \in \mathbb{R}^{n\times k}$ とベクトルを結合して表記すると, $\widehat{\boldsymbol{\beta}}_n$ は $\mathbf{X}^\top\mathbf{X}$ が正則ならば以下で与えられる:

$$\widehat{\boldsymbol{\beta}}_n = \left(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^\top\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_iY_i\right) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$$

OLSE を構成するために用いる目的関数は, MSE ではないことが分かった. 最終的には本書でも, 定義として採用されたのは後者の方法.

■$\mathbb{E}[\mathbf{X}\mathbf{X}^\top]$ が正則ならば正定値　線形回帰モデルにおける $\boldsymbol{\beta}$ の解釈を与える, **Thm 9.11** において, SOSC の確認のため利用されている事実だが, 特に明示して証明されていなかったため, ここで示しておく.

Thm

$\mathbf{X} \in \mathbb{R}^k$ とするとき, $\mathbb{E}[\mathbf{X}\mathbf{X}^\top] \in \mathbb{R}^{k\times k}$ が正則であるならば正定値である.

Proof 任意の非零ベクトル $\mathbf{v} \in \mathbb{R}^k$ に対して

$$\mathbf{v}^\top A\,\mathbf{v} = \mathbf{v}^\top\,\mathbb{E}[\mathbf{X}\mathbf{X}^\top]\,\mathbf{v} = \mathbb{E}\big[\mathbf{v}^\top(\mathbf{X}\mathbf{X}^\top)\mathbf{v}\big] = \mathbb{E}\big[(\mathbf{v}^\top\mathbf{X})^2\big] \;\geq\; 0$$

が成り立つので, $\mathbf{A}$ は半正定値である. A は対称かつ半正定値であるので, 直交行列 $\mathbf{Q}$ を用いて

$$\mathbf{Q}^\top\mathbf{A}\,\mathbf{Q} = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$$

と対角化でき, 半正定値性から各固有値 λ_i は $\lambda_i \geq 0$ である. ここで正則性 $\det \mathbf{A} \neq 0$ から

$$\det \mathbf{A} = \prod_{i=1}^{k} \lambda_i \neq 0$$

より, 全ての i に対して $\lambda_i > 0$ が得られる. 以上より, 任意の非零ベクトル $\mathbf{v} \in \mathbb{R}^k$ に対して

$$\mathbf{v}^\top \mathbf{A}\, \mathbf{v} = \mathbf{v}^\top \mathbf{Q}\, \mathrm{diag}(\lambda_1, \ldots, \lambda_k)\, \mathbf{Q}^\top \mathbf{v} = \sum_{i=1}^{k} \lambda_i\, w_i^2 > 0$$

（ただし $\mathbf{w} = \mathbf{Q}^\top \mathbf{v} \neq \mathbf{0}$）が成り立つので, $\mathbf{A}$ は正定値である. □

一般に, 正定値性は正則性を意味する. だが, 今回の例で示された逆の関係 (正則ならば正定値行列) は, 一般の正則な対称行列には成り立たない関係であることを注意せよ. 半正定値かつ正則ならば, 正定値である, これに関しては証明後半と同様の議論によって, 一般に同値関係が成り立つ. 例えばここを参照せよ: **証明**

Thm: 半正定値行列と正則性

半正定値行列が正定値行列であるための必要十分条件は, 正則なことである.

これを利用して, 以下のように完全な多重共線性 (フルランク性) の排除の意味を考えることが出来る.

Thm: $\mathbf{X}^\top\mathbf{X}$ に課される正則性の意味

$\mathbf{X} \in \mathbb{R}^{n\times k}$ とするとき, $\mathbf{X}^\top\mathbf{X} \in \mathbb{R}^{k\times k}$ が正則であるならば正定値である.

Proof まず, $\mathbf{X}^\top\mathbf{X}$ が半正定値であることを示す. 任意の非ゼロベクトル $\mathbf{v} \in \mathbb{R}^k$ に対して,

$$\mathbf{v}^\top \mathbf{X}^\top \mathbf{X}\mathbf{v} = (\mathbf{X}\mathbf{v})^\top (\mathbf{X}\mathbf{v}) = \|\mathbf{X}\mathbf{v}\|^2 \geq 0$$

が成り立つため, $\mathbf{X}\mathbf{X}^\top$ は半正定値. さらに上の定理と, 正則性の仮定より題意は満たされた. □

Thm 6.8: 二次形式の最小化における十分条件 (SOSC)

関数 $f : \mathbb{R}^k \to \mathbb{R}$:

$$f(\mathbf{b}) = \frac{1}{2}\mathbf{b}^\top \mathbf{A}\mathbf{b} + \mathbf{a}^\top \mathbf{b} + c$$

で, $\mathbf{A}$ は所与の対称行列, $\mathbf{a}$ は定数ベクトル, c は定数とする. このとき, 一階条件の $\mathbf{D}f(\mathbf{b}) = \mathbf{0}$ を満たす $\widehat{\mathbf{b}}$ が $f(\mathbf{b})$ を最小化するための十分条件は, $\mathbf{A}$ が半正定値なこと. 二階の十分条件 (SOSC) という.

計量経済学で登場する大概の最適化問題の十分条件として, 上の定理から半正定値性を確認すれば十分. 実際, OLS や GLS に関しては, 一階条件が実際に目的関数を最小化することは, この確認さえすれば十分である.

■OLSE や GLSE と正定値性　OLSE や GLSE を考える際に, $\mathbf{A}$ にあたる部分が半正定値になるように構成されているわけだが, この部分は推定量を表す際に逆行列をとられる. これは正則性が必要な操作であり, 従ってこの部分は各推定量の定義をした時点で, 正定値であることが確定する.

Def 6.2: マハラノビス距離

ベクトル $\mathbf{b} \in \mathbb{R}^k$ と, k 次対称 (半) 正定値行列 $\mathbf{A}$ に対し, $\mathbf{A}$ をウェイトとするノルムを:

$$\|\mathbf{b}\|_{\mathbf{A}} \coloneqq \sqrt{\mathbf{b}^\top \mathbf{A}\mathbf{b}}$$

と定義し, ベクトル $\mathbf{b}_1, \mathbf{b}_2 \in \mathbb{R}^k$ に対して, マハラノビス距離 $d_{\mathbf{A}}(\mathbf{b}_1, \mathbf{b}_2)$ は次で定義される:

$$d_{\mathbf{A}}(\mathbf{b}_1, \mathbf{b}_2) \coloneqq \|\mathbf{b}_1 - \mathbf{b}_2\|_{\mathbf{A}}$$

Def: GLS の目的関数

正定値行列 $\mathbf{W} \in \mathbb{R}^{n \times n}$ に対し, 任意のベクトル $\mathbf{e} \in \mathbb{R}^n$ に対する一般化ノルムは次で定義される:

$$\|\mathbf{e}\|_{\mathbf{W}} \coloneqq \sqrt{\mathbf{e}^\top \mathbf{W}\mathbf{e}}$$

GLS では, 残差ベクトル $\mathbf{e} = \mathbf{Y} - \mathbf{X}\mathbf{b}$ のこの一般化ノルムを最小化することによって推定値を求める.

重み付け行列が正定値であることについては, 末石計量の補足資料, GLS の効率性の章で触れた. 重み付け行列として不均一分散の下で, 共分散行列の逆行列を利用するのが一般的だが, これは正則性を課していることと同値で, 共分散行列の半正定値性から, 正定値性を仮定していることに他ならない. 正定値の逆行列も正定値であるため, この一般形にあたる $\mathbf{W}$ は初めから正定値として仮定されているのだと考えられる. なお, 『計量経済学のための数学』においても, 重み付け行列は対角行列に直ちに限定されて GLSE は定義される.

■GLSE の一致性の条件　GLS の一致性の条件からも得られる含意がある. 確認してみよう.

Thm 10.8, 11.6: GLS 推定量の条件

線形回帰モデル $\mathbb{E}[Y \mid \mathbf{X}] = \mathbf{X}^\top \boldsymbol{\beta}$ について, 以下の (i)〜(iii) が満たされるとする:

1. $\mathbb{E}[\mathbf{X}^\top \mathbf{X}] < \infty, \quad \mathbb{E}[Y^2] < \infty,$
2. W は $\mathbf{X}$ 可測, $\quad \mathbf{P}\{W > 0\} = 1, \mathbb{E}[W^2 \mathbf{X}^\top \mathbf{X}] < \infty, \mathbb{E}[\varepsilon^2 W^2 \mathbf{X}^\top \mathbf{X}] < \infty,$
3. $\mathbf{X}_n^\top W_n \mathbf{X}_n, \mathbb{E}[W \mathbf{X}\mathbf{X}^\top]$ は正則

このとき, GLS 推定量について一致性 $\widehat{\boldsymbol{\beta}}_n^w \xrightarrow{p} \boldsymbol{\beta}$ と漸近正規性が成立する.

期待値の有限性は期待値の線形性を経由して LLN に利用される. W の $\mathbf{X}$ 可測は LIE で利用されるが, 具体例として, 不均一分散の場合も当てはまる. 11 章で後述されるが, これは, (条件付の) 共分散行列が対角行列となり, その対角成分が $\mathbf{X}$ の関数 $\sigma(\cdot)$ として表される状況である. **Thm 9.4** での議論から, これは $\mathbf{X}$ 可測である. 正則性は先述の通りで, GLSE の表現に必要となる.

Thm 11.7: 一般化コーシー・シュワルツの不等式

任意の確率変数ベクトル $\mathbf{X} \in \mathbb{R}^k$, $\mathbf{Y} \in \mathbb{R}^l$ について, $\mathbb{E}[\mathbf{X}\mathbf{X}^\top], \mathbb{E}[\mathbf{Y}\mathbf{Y}^\top]$ が正則なとき, 以下が成立する:

$$\mathbb{E}[\mathbf{Y}\mathbf{X}^\top][\mathbb{E}[\mathbf{X}\mathbf{X}^\top]]^{-1}\mathbb{E}[\mathbf{X}\mathbf{Y}^\top] \leq \mathbb{E}[\mathbf{Y}\mathbf{Y}^\top]$$

ここで, $\mathbf{A} \leq \mathbf{B}$ は差分 $\mathbf{B} - \mathbf{A}$ が半正定値であることを意味する.

この不等式を利用して, GLSE における効率的な重み付け行列を特定化できる.

Thm 11.8: GLS 推定量の最適な重み付け行列

Thm 11.6 を満たす任意の確率変数 W について GLSE の漸近分散を $\mathbf{V}(W)$ として, 以下が成り立つ:

$$\mathbf{V}(\frac{1}{\sigma_\varepsilon^2(\mathbf{X})}) = \big[\mathbb{E}[\frac{\mathbf{X}\mathbf{X}^\top}{\sigma_\varepsilon^2(\mathbf{X})}]\big]^{-1} \leq \mathbf{V}(W)$$

つまり, 条件付分散 $\sigma_\varepsilon^2(\mathbf{X}_i) = \mathbb{E}[\varepsilon^2 \mid \mathbf{X}_i]$ を対角成分にもつ対角行列の逆数, 若しくはこの逆数を対角成分にもつ対角行列が最良の重み付け行列となる.

■誤差項の i.i.d. **性** (p.224)　当然視される気もするがあまり自明とは思えない事実. 実際, 自分は独立性のみが成立すると認識していた. ここでは, **付録 C** の測度論的確率論の諸定義, 諸定理を活用しながら証明を行う.

Thm: 誤差項の i.i.d. 性

確率変数ベクトル列 $\{\mathbf{Z}\} = \{[Y, \mathbf{X}^\top]^\top\}$ が独立同一であるならば, $\varepsilon = Y - \mathbb{E}[Y \mid \mathbf{X}]$ によって与えられる $\{\boldsymbol{\varepsilon}\}$ も独立同一に分布する確率変数になる.

Proof まず, 各標本 $i \in \{1, 2, \dots\}$ に対して $\mathbf{Z}_i$ の実現値 (データ) $\mathbf{z}_i = (y_i, \mathbf{x}_i)$ が得られた場合:

$$\varepsilon_i = g(\mathbf{Z}_i), \quad g(\mathbf{Z}_i) \coloneqq Y_i - \mathbb{E}[\,Y_i \mid \mathbf{X}_i\,], \\ g(\mathbf{Z}_i = \mathbf{z}_i) = y_i - \mathbb{E}[\,Y_i \mid \mathbf{X}_i = \mathbf{x}_i\,]$$

と定義できる. ここで, ε_i もまた確率変数である. このとき, **Def 9.5** より条件付期待値である第二項の $\mathbf{X}_i$ 可測性が言えて, $\mathbf{Z}_i$ 可測であることも分かる. $g(\mathbf{Z}_i)$ は $\mathbf{Z}_i$ 可測. 実際,

$$\mathbb{E}[g(\mathbf{Z}_i) \mid \mathbf{Z}_i] = \mathbb{E}[Y_i - \mathbb{E}[\,Y_i \mid \mathbf{X}_i\,] \mid \mathbf{Z}_i] = \mathbb{E}[Y_i - \mathbb{E}[\,Y_i \mid \mathbf{X}_i\,] \mid Y_i, \mathbf{X}_i] = Y_i - \mathbb{E}[\,Y_i \mid \mathbf{X}_i\,] = g(\mathbf{Z}_i)$$

の関係が, **Thm 9.8** から成り立つことが確認できる. **付録 C.1** の $\mathbf{Z}_i$ 可測の定義から, 確率変数によって生成される σ-加法族の関係性として以下が成立している:

$$\sigma[\varepsilon_i] \ \subset \ \sigma[\mathbf{Z}_i]$$

(1) 独立性.　任意の $n \in \mathbb{N}$ に対して $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ が独立であるとは, **Def 10.2.2** より:

$$\sigma[\mathbf{Z}_1], \dots, \sigma[\mathbf{Z}_n] \quad \text{が互いに独立}$$

であることをいう. 各 $\sigma[\varepsilon_i]$ は $\sigma[\mathbf{Z}_i]$ の部分集合族であるので, 集合族の独立性 **Def 10.2.1** から, 集合族 $\mathcal{G}_i \coloneqq \sigma[\varepsilon_i] \subset \sigma[\mathbf{Z}_i]$ から選ばれた任意の事象 $A_i \in \mathcal{G}_i \subset \sigma[\mathbf{Z}_i]$ も常に独立である. よって:

$$\sigma[\varepsilon_1], \dots, \sigma[\varepsilon_n] \quad \text{も互いに独立}$$

となり, すなわち任意の $n \in \mathbb{N}$ に対して $\varepsilon_1, \dots, \varepsilon_n$ は互いに独立である.

(2) 同一分布性.　$\{\mathbf{Z}_i\}$ が同一分布とは, ある分布関数 $F_{\mathbf{Z}}(\cdot)$ が存在して, 全ての $\mathbf{t} \in \mathbb{R}^k$ について

$$\mathbf{P}\{\mathbf{Z}_i \leq \mathbf{t}\} = F_{\mathbf{Z}}(\mathbf{t}) \quad (i = 1, 2, \dots)$$

であること. $\varepsilon_i = g(\mathbf{Z}_i)$ は同一の可測写像 g による変換であるから, 全ての $\tau \in \mathbb{R}$ に対して

$$\mathbf{P}\{\varepsilon_i \leq \tau\} = \mathbf{P}\{\,g(\mathbf{Z}_i) \leq \tau\} = F_\varepsilon(\tau) \quad (i = 1, 2, \dots)$$

となり, $\{\varepsilon_i\}$ は同一分布である.

以上 (1)(2) より, **付録 C.4** の定義と整合し, $\{\varepsilon_i\}$ が i.i.d. であることが示された. □

B.2　線形代数に関連する諸注意

『計量経済学のための数学』では扱われていない, 完全な多重共線性の連立方程式の解システムとしての問題点等, 連立方程式システムに注目した際の線形代数の諸定理を確認する.

■ランク $\leq$ 行/列数　$n \times k$ 行列 $\mathbf{A}$ について, $\mathrm{rank}(\mathbf{A}) \leq \min\{n, k\}$ が成り立つ **証明**

■列フルランクの必要条件　$n \times k$ 行列 $\mathbf{A}$ が列フルランクである必要条件は $n \geq k$ である

■行列積と rank　$\mathrm{rank}(\mathbf{AB}) \leq \min\{\mathrm{rank}(\mathbf{A}), \mathrm{rank}(\mathbf{B})\}$ が任意の 2 行列で成立する **証明**

■正則行列と行列の積での rank　行列の積が定義される任意の正則行列 $\mathbf{B}, \mathbf{C}$ と任意の行列 $\mathbf{A}$ について, $\mathrm{rank}(\mathbf{AB}) = \mathrm{rank}(\mathbf{A})$, $\mathrm{rank}(\mathbf{AC}) = \mathrm{rank}(\mathbf{A})$ が成立する **証明**

■正則性と同値な概念　$\mathbf{A}$ が正則行列 $\iff$ 逆行列が存在する $\iff$ $\det(\mathbf{A}) \neq 0$ $\iff$ 列ベクトルの線形独立 $\iff$ フルランク $\iff$ $\mathbf{Ax} = \mathbf{b}$ の解 $\mathbf{x}$ が一意 $\iff$ $\mathbf{Ax} = \mathbf{0}$ の解が, 自明な解 $\mathbf{x} = \mathbf{0}$ のみ **証明**

■正則と正定値行列　正定値行列は正則で逆行列も正定値行列 **証明**

■フルランク行列と行列の積での rank　行列の積が定義される任意の列フルランク行列 $\mathbf{B}$, 行フルランク行列 $\mathbf{C}$ と任意の行列 $\mathbf{A}$ について, $\mathrm{rank}(\mathbf{BA}) = \mathrm{rank}(\mathbf{A})$, $\mathrm{rank}(\mathbf{AC}) = \mathrm{rank}(\mathbf{A})$ が成立する. 証明不明.

■行列 $\mathbf{X}^\top\mathbf{X}$ の性質　任意の $n \times k$ 行列 $\mathbf{X}$ について, $\mathbf{X}^\top\mathbf{X}$ は定義され, k 次正方行列となる. また, 必ず対称行列となる. うえで見た性質を利用すれば, ランクについては以下が成り立つことを確認できる:

$$\mathrm{rank}(\mathbf{X}^\top\mathbf{X}) \leq \min\{\mathrm{rank}(\mathbf{X}^\top), \mathrm{rank}(\mathbf{X})\} = \mathrm{rank}(\mathbf{X}) \leq \min\{n, k\}$$

$n < k$ のとき, 上の不等式の関係から $\mathrm{rank}(\mathbf{X}^\top\mathbf{X}) < k$ が成り立ち, $\mathbf{X}^\top\mathbf{X}$ は正則になり得ない. つまり, $n \geq k$ は $\mathbf{X}^\top\mathbf{X}$ が正則であるための必要条件なことが分かる. **$\mathbf{X}^\top\mathbf{X}$ の半正定値性** も確認したとおり. また, $\mathrm{rank}(\mathbf{X}) < k$ なら $\mathrm{rank}(\mathbf{X}^\top\mathbf{X}) < k$ が成り立ち, 同様に $\mathbf{X}$ の列フルランク性も正則性の必要条件である.

Thm: 行列 $\mathbf{X}^\top\mathbf{X}$ の性質

任意の $n \times k$ 行列 $\mathbf{X}$ からなる k 次対称正方行列 $\mathbf{X}^\top\mathbf{X}$ は以下の性質を持つ:

1. 半正定値行列
2. 正則ならば正定値行列
3. $\mathbf{X}$ の列フルランク性, $n \geq k$ は $\mathbf{X}^\top\mathbf{X}$ の正則性の必要条件

付録 C　測度論的確率論

末石計量においても測度論的確率論についての記載は付録にもなく, i.i.d. 性の議論等が困難になっている. いくつかの定義については議論において利用することを避けて通れないため, 利用できるものを確認しておこう. 詳しくは『計量経済学のための数学』の 7 ~ 11 章を参照のこと. 定義・定理番号は教科書に対応する.

C.1 確率空間 (7 章)

Def 7.1: σ-加法族

集合 Ω 上の部分集合族 $\mathcal{F} \subset 2^{\Omega}$ が, 次の 3 条件を満たすとき, $\mathcal{F}$ を σ-加法族という.

1. $\Omega \in \mathcal{F}$.
2. $A \in \mathcal{F}$ ならば $A^c \coloneqq \Omega \setminus A \in \mathcal{F}$.
3. 任意の列 $\{A_n\}_{n=1}^{\infty} \subset \mathcal{F}$ に対して, $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

Def 7.4: σ-加法族の生成

$\mathcal{F}_0$ を含む最小の σ-加法族 $\mathcal{F}$ を, $\mathcal{F}_0$ から生成された σ-加法族と呼び, 以下のように表す:

$$\mathcal{F} \coloneqq \sigma\big[\mathcal{F}_0\big]$$

Def 7.5: ボレル集合族

$\mathbb{R}^k$ 上の, あらゆる直方体 $\Pi_{i=1}^{k}(a_i, b_i)$ からなる集合族 $\mathcal{G}$ より生成された (最小の) σ-加法族を, (k 次元) ボレル集合族といい, 以下のように表記される:

$$\mathcal{B} \coloneqq \sigma\big[\{\Pi_{i=1}^{k}(a_i, b_i) \mid a_i < b_i \, \forall i\}\big] \, .$$

ボレル集合族は標本空間が $\mathbb{R}^k$ の場合には事象族 $\mathcal{F}$ として利用する.

Def: 可測空間

標本空間 (確率空間の諸概念構成に用いられる全体集合) Ω とその上の σ-加法族である事象族 $\mathcal{F}$ の組

$$(\Omega, \mathcal{F})$$

を可測空間という. また, $A \in \mathcal{F}$ である事象 A を可測集合という.

Def 7.2: 確率測度

可測空間 $(\Omega, \mathcal{F})$ 上の写像 $\mathbf{P} : \mathcal{F} \to [0, 1]$ が以下を満たすとき, $\mathbf{P}$ を確率測度という.

1. $\mathbf{P}\Omega = 1$
2. $\mathbf{P}(A^C) = 1 - \mathbf{P}A$
3. 互いに素な列 $\{A_n\}_{n=1}^{\infty} \subset \mathcal{F}$ に対して $\mathbf{P}\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \sum_{n=1}^{\infty} \mathbf{P}A_n$

Def: 確率空間

標本空間 Ω とその上の事象族 $\mathcal{F}$, およびこの可測空間 $(\Omega, \mathcal{F})$ 上の確率測度 $\mathbf{P}$ の組

$$(\Omega, \mathcal{F}, \mathbf{P})$$

を確率空間という.

Def 7.3: 確率質量関数

加算集合で表現される標本空間 $\Omega = \{\omega_1, \omega_2, \ldots\}$ について, 事象族としてべき集合族 2^Ω を選択した確率空間 $(\Omega, 2^\Omega, \mathbf{P})$ 上で, 各 $\omega \in \Omega$ に対して

$$p(\omega) \coloneqq \mathbf{P}\{\omega\}$$

で定義される関数 $p : \Omega \to [0, 1]$ を $\mathbf{P}$ の確率質量関数という.

Def 7.6: 確率変数

可測空間 $(\Omega, \mathcal{F})$ 上の写像 (関数) $X : \Omega \to \mathbb{R}$ が任意の $t \in \mathbb{R}$ について

$$\{\omega \in \Omega \mid X(\omega) \leq t\} \in \mathcal{F}$$

を満たすとき, X を確率変数という.

確率変数の定義自体には確率測度は介在しない.

Def: 確率変数ベクトル

可測空間 $(\Omega, \mathcal{F})$ 上の写像 (ベクトル値関数) $\mathbf{X} : \Omega \to \mathbb{R}^k$ が任意のベクトル $\mathbf{t} \in \mathbb{R}^k$ について

$$\{\omega \in \Omega \mid \mathbf{X}(\omega) \leq \mathbf{t}\} \in \mathcal{F}$$

を満たすとき, $\mathbf{X}$ を確率変数ベクトルという.

確率変数は単に関数のこと. 『扱いにくい不確実性は全て標本空間に押し付けられている』.

Notation: 事象に関する略記

「関数 X が t 以下の値をとる」ような事象 ($\times$ 状態 ω) $\{\omega \in \Omega \mid X(\omega) \leq t\} \in \mathcal{F}$ を以下のように記す:

$$\{X \leq t\}.$$

「関数 X のとる値が D に属する」ような事象 $\{\omega \in \Omega \mid X(\omega) \in D\} = X^{-1}(D)$ を以下のように記す:

$$\{X \in D\}.$$

X のある実現値 x が判明したとき, 確率変数の関係を辿れば事象 $\{X \leq t\}$ where $t \in \mathbb{R}$ の成立/不成立が (すべてではないが) 判明する. ここで成立/不成立が判明した事象を集めれば σ-加法族が生成される.

Def: 確率変数が生成する事象族

$$\sigma[X] \coloneqq \sigma\big[\{\{X \leq t\} \mid t \in \mathbb{R}\}\big]$$

値域が加算集合の離散確率変数については, 以下のように表現が可能.

Thm 7.2: 離散確率変数が生成する事象族

X の値域 $\mathcal{X} = X(\Omega)$ が加算集合であるとき, 以下が成り立つ:

$$\sigma[X] = \{\{X \in \mathcal{X}_0\} \mid \mathcal{X}_0 \subset \mathcal{X}\}$$

証明前半は σ-加法族であることを確認しているが, 後半は何をやっている?

Def: X 可測な確率変数

同じ可測空間上で与えられた確率変数 X, Y について以下が成り立つとき, Y は X 可測であるという.

$$\sigma\big[Y\big] \subset \sigma\big[X\big]$$

直感的には, 『X の値が観測できれば Y の値も分かる』, Y の情報は X より劣っている. 条件付期待値やそれに付随する LIE で登場する情報集合の包含関係に関する議論はここにたどり着く.

Def 7.7 & Thm 7.3, 7.4: 分布関数

確率空間 $(\Omega, \mathcal{F}, \mathbf{P})$ 上の確率変数 $X : \Omega \to \mathbb{R}$ に対し, その分布関数 $F_X : \mathbb{R} \to [0,1]$ を各 $t \in \mathbb{R}$ で:

$$F_X(t) \coloneqq \mathbf{P}\big\{\omega \in \Omega \mid X(\omega) \leq t\big\} = \mathbf{P}\{X \leq t\}$$

で定義する. この関数 F_X は次の性質を満たす:

1. 非減少性: $x \leq y \implies F_X(x) \leq F_X(y)$.
2. 境界条件: $\lim_{x\to-\infty} F_X(x) = 0,\ \lim_{x\to+\infty} F_X(x) = 1$.

確率変数の離散/連続は分布関数の連続性によって区別される.

Def 7.8: 密度関数

分布関数 F_X を持つ連続確率変数 $X : \Omega \to \mathbb{R}$ について, 任意の $t \in \mathbb{R}$ に対して

$$F_X(t) = \int_{-\infty}^{t} f_X(x)\, dx, \qquad f_X(x) = \frac{dF_X}{dx}(x)$$

が成り立つ可積分関数 $f_X : \mathbb{R} \to \mathbb{R}$ が存在するとき, f_X を密度関数という. f_X は次の性質を満たす:

1. 非負性: $f_X(x) \geq 0 \ \ \forall x \in \mathbb{R}$
2. 正規化条件: $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$

Def 7.9: 確率変数ベクトルの密度関数

分布関数 $F_{\mathbf{X}}(\mathbf{t}) = \mathbf{P}\{\mathbf{X} \leq \mathbf{t}\}$ を持つ連続な $\mathbf{X} : \Omega \to \mathbb{R}^k$ を考え, 任意の $t \in \mathbb{R}^k$ に対して

$$F_{\mathbf{X}}(\mathbf{t}) = \int_{-\infty}^{t_1} \cdots \int_{-\infty}^{t_k} f_{\mathbf{X}}(\mathbf{x})\, dx_1 \cdots dx_k, \qquad f_{\mathbf{X}}(\mathbf{x}) = \frac{\partial^k F_{\mathbf{X}}}{\partial x_1 \partial x_2 \cdots \partial x_k}(\mathbf{x})$$

が成り立つ可積分関数 $f_{\mathbf{X}} : \mathbb{R}^k \to \mathbb{R}$ が存在するとき, $f_{\mathbf{X}}$ を密度関数という.

密度関数から確率測度を構成することも可能.

Def 7.10: 確率変数の質量関数

加算集合で表現される実現値 $\mathcal{X} = \{x_1, x_2, \ldots\}$ を持つ離散確率変数 X について,

$$p_X(x_j) \coloneqq \mathbf{P}\{X = x_j\}$$

で定義される関数 $p_X : \mathcal{X} \to [0,1]$ を確率変数 X の確率質量関数という. p_X は次の性質を満たす:

1. 非負性: $p_X(x) \geq 0 \ \ \forall x \in \mathcal{X}$
2. 正規化条件: $\sum_{j=1}^{\infty} p_X(x_j) = 1$

確率測度に対して定義したものとの差異は定義域程度か. 確率質量関数から確率測度を構成することも可能.

C.2 積分と期待値 (8 章)

■指示関数　最も単純な構造を持つ確率変数として解釈できる. 『計量経済学のための数学』においては, (8.5) 式での指示関数で表現できる X と例題 7.8 で定義された Y は同じ確率関数で, 従ってこれらの確率変数から生成される σ-加法族なども同じであることに気付くだろう.

■単関数　確率変数の集合が線形空間になる事から, 指示関数の線形結合で表せる単関数も確率変数である.

Def 8.1: 単関数

確率空間 $(\Omega, \mathcal{F}, \mathbf{P})$ に対し, 関数 $X : \Omega \to \mathbb{R}$ が次の形で表されるとき, X を単関数という:

$$X(\omega) \coloneqq \sum_{i=1}^{n} c_i \mathbb{1}_{A_i}(\omega)$$

ここで, $c_i \in \mathbb{R}$, $A_i \in \mathcal{F}$ は互いに素な集合, $\mathbb{1}_{A_i}$ は A_i の指示関数 ($\mathbb{1}_{A_i}(\omega) = 1$ if $\omega \in A_i$, o.w. 0).

Def 8.2: 単関数の期待値

確率空間 $(\Omega, \mathcal{F}, \mathbf{P})$ 上の単関数 $X(\omega)$ に対して, その期待値 $\mathbb{E}[X]$ を次で定義する:

$$\mathbb{E}[X] \coloneqq \sum_{i=1}^{n} c_i \mathbf{P} A_i$$

Def 8.3: 近似単関数

確率空間 $(\Omega, \mathcal{F}, \mathbf{P})$ 上の非負かつ有限な確率変数 $X : \Omega \to [0, M]$ に対し以下で定義される:

$$X_n \coloneqq \sum_{i=1}^{2^n} \frac{i-1}{2^n} M \mathbb{1}_{A_{n,i}}$$

X_n を, 近似単関数と呼ぶ. ここで, $A_{n,i}$ は X の値域を 2^n 等分し, このうち小さいほうから j 番目の小区間に X が含まれる事象を指す. $\{X_n\}_{i=1}^{\infty}$ は上に有界かつ単調増加な数列であるため以下が成り立つ:

$$\lim_{n \to \infty} X_n(\omega) = X(\omega)$$

一般の確率変数では単関数による近似等多様な操作を駆使し, 単関数の期待値の定義に帰着させる.

Thm 8.2: 離散確率変数の期待値

加算集合で表現される実現値 $\mathcal{X} = \{x_1, x_2, \ldots\}$ をとる確率変数 $X : \Omega \to \mathcal{X}$ に対し, 確率質量関数 $p_X(x_j) = \mathbf{P}\{X = x_j\}$ を用いて, 期待値 $\mathbb{E}[\,X\,]$ は次の通り:

$$\mathbb{E}[\,X\,] \coloneqq \sum_{j=1}^{\infty} x_j \, p_X\,(x_j)$$

Thm 8.3: 連続確率変数の期待値

確率密度関数 f_X をもつ連続確率変数 $X : \Omega \to \mathcal{X} \subset \mathbb{R}$ に, 期待値 $\mathbb{E}[\,X\,]$ は次の通り:

$$\mathbb{E}[\,X\,] \coloneqq \int_{\mathcal{X}} x \, f_X\,(x)\, dx$$

確率密度関数の存在と, 分布関数の絶対連続性を前提として, スティルチェス積分の利用により得られる.

C.3 条件付期待値と回帰分析 (9 章)

Def: 結合確率質量関数

離散確率変数ベクトル $\mathbf{Z} = (Y, \mathbf{X})$ が可算集合 $\mathcal{Z} = \{(y_1, \mathbf{x}_1), \ldots\}$ から実現値を取るとき, 結合確率質量関数 $p_{Y,\mathbf{X}} : \mathcal{Z} \to [0,1]$ を各 $(y, \mathbf{x}) \in \mathcal{Z}$ に対して以下で定義する:

$$p_{Y,\mathbf{X}}(y, \mathbf{x}) \coloneqq \mathbf{P}\{Y = y, \mathbf{X} = \mathbf{x}\}$$

まずは加算集合から実現値をとる場合の, 条件付確率にまつわる諸概念を考えていこう.

Def: 周辺確率質量関数

結合分布と同様の状況で周辺確率質量関数を以下で定義する:

$$p_Y(y) \coloneqq \mathbf{P}\{Y = y\}, \quad p_{\mathbf{X}}(\mathbf{x}) \coloneqq \mathbf{P}\{\mathbf{X} = \mathbf{x}\}$$

Thm 9.1: 結合分布と周辺分布の関係

$\mathcal{Y} = \{y_1, y_2, \ldots\}$, $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$ を値域として以下が成り立つ:

$$p_Y(y) = \sum_{\mathbf{x} \in \mathcal{X}} p_{Y,\mathbf{X}}(y, \mathbf{x}), \quad p_{\mathbf{X}}(\mathbf{x}) = \sum_{y \in \mathcal{Y}} p_{Y,\mathbf{X}}(y, \mathbf{x}).$$

確率測度の σ-加法性から成立する.

Def 9.1: 条件付確率質量関数

条件付確率質量関数 $p_{Y|\mathbf{X}} : \mathcal{Z} \to [0,1]$ を以下で定義する:

$$p_{Y|\mathbf{X}}(y \mid \mathbf{x}) \coloneqq \frac{p_{Y,\mathbf{X}}(y, \mathbf{x})}{p_{\mathbf{X}}(\mathbf{x})}$$

Def 9.2: 離散確率変数の条件付期待値

$\mathbf{X} = \mathbf{x}$ に条件付けられた, $Y = y$ の, $\mathcal{X}$ 上の条件付期待値を以下で定義する:

$$\mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}] \coloneqq \sum_{y \in \mathcal{Y}} y\, p_{Y|\mathbf{X}}(y \mid \mathbf{x})$$

■条件付期待値は $\mathcal{X}$ 上の関数　なぜこのように言えるのだろうか. 条件付 pmf は実数値をとるため, とる値は定義を見る限りでは実数値に過ぎない. 確認してみよう. $\mathbf{X} = \mathbf{x} \in \mathcal{X}$ に条件付けられた条件付期待値を以下のように $\mathbf{x}$ の関数 ϕ として表現する:

$$\phi(\mathbf{x}) \coloneqq \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$$

ここで, $\mathbf{x}$ は $\mathcal{X}$ 上の実現値をとることが分かっているため, $\phi : \mathcal{X} \to \mathbb{R}$ である. これは $\mathcal{X}$ 上の関数である. 自身の誤解は, 『$\mathcal{X}$ 上の関数』という言葉を, 値域が $\mathcal{X}$ である, と誤解したことによるものだった.

Notation: 条件付期待値の略記

実現値 $\mathbf{x}$ を特定しない場合は以下のように記す:

$$\mathbb{E}[Y \mid \mathbf{X}] \,.$$

これは各 $\omega \in \Omega$ に対して $\mathbb{E}[Y \mid \mathbf{X} = \mathbf{X}(\omega)]$ を対応させる, $\Omega \to \mathbb{R}$ の関数, 即ち**確率変数**と見なせる.

Thm 9.2: 繰り返し期待値の法則 (LIE)

加算集合 $\mathcal{Y} = \{y_1, y_2, \ldots\}$, $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$ 上の離散確率変数 $Y, \mathbf{X}$ に対して以下が成り立つ:

$$\mathbb{E}[\mathbb{E}[Y \mid \mathbf{X}]] = \mathbb{E}[Y] \,.$$

証明では外側の期待値について, 条件付期待値を確率変数と見なせることによって, **Thm 8.2** を使用できる. 更に, 各実現値を考えるにあたっては条件付期待値は $\mathcal{X}$ 上の関数としても考えられたため, 証明 (p.178) の一行目の変形が得られる. 後は明らか.

Thm 9.3: 条件付期待値の性質

加算集合 $\mathcal{Y} = \{y_1, y_2, \ldots\}$, $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$ 上の離散確率変数 $Y, \mathbf{X}$, 任意の関数 $g(\mathbf{X})$ に対して:

$$\mathbb{E}[g(\mathbf{X})Y \mid \mathbf{X}] = g(\mathbf{X})\mathbb{E}[Y \mid \mathbf{X}] \,.$$

$\mathcal{Y}_0 \subset \mathcal{Y}$ と Y の取る値を限定している部分はテクニカル.

Thm 9.4: 条件付期待値と X 可測性

条件付期待値 $\mathbb{E}[Y \mid \mathbf{X}]$ は $\mathbf{X}$ 可測である. さらに, 任意の $\mathbf{X}$ 可測集合 $A \in \sigma[\mathbf{X}]$ で以下が成り立つ:

$$\mathbb{E}[\mathbb{1}_A\, \mathbb{E}[Y \mid \mathbf{X}]] = \mathbb{E}[\mathbb{1}_A\, Y] \,.$$

証明には教科書で触れられていない行間が存在する. 任意の $\mathbf{X}$ の関数 Y が $\mathbf{X}$ 可測であることを示さねばならないはずだが, 教科書では特に記載がない. 問題 7.6 でも特定の関数について可測性を示すにとどまっている. 後半の証明は **Thm 7.2**, **Thm 9.3**, **Thm 9.2** による.

■一般の条件付期待値　いよいよ離散の場合に限らない, 一般の条件付期待値を考える.

Def 9.3: 部分 σ-加法族

σ-加法族 $\mathcal{F}$ の部分集合族 $\mathcal{G} \subset \mathcal{F}$ が, σ-加法族の条件を満たすとき, $\mathcal{G}$ を $\mathcal{F}$ の部分 σ-加法族という.

Def 9.4: 一般の条件付期待値

確率空間 $(\Omega, \mathcal{F}, \mathbf{P})$ 上の確率変数 Y と, $\mathcal{F}$ の部分 σ-加法族 $\mathcal{G} \subset \mathcal{F}$ に対して, 以下の性質を持つ確率変数 W を, Y の $\mathcal{G}$ に条件付けられた期待値といい, $W = \mathbb{E}[Y \mid \mathcal{G}]$ と書く:

1. $\mathbb{E}[X \mid \mathcal{G}]$ は $\mathcal{G}$-可測, 即ち任意の実数 t について $\{W \leq t\} \in \mathcal{G}$
2. 任意の $A \in \mathcal{G}$ に対して以下が成立する:

$$\mathbb{E}[\mathbb{1}_A W] = \mathbb{E}[\mathbb{1}_A Y]$$

とくに, $A = \Omega$ とするとき $\mathbb{E}[W] = \mathbb{E}[Y]$

$\mathcal{G}$-可測はここで初めて定義された言葉. **Thm 9.4** の一般化を定義として議論が始まっている.

■部分 σ-加法族の意味　記述を眺めてみよう. いい解説.

田中 久稔, 計量経済学のための数学, p.183

『事象族 $\mathcal{F}$ には「試行の結果を観測することによって明らかになる情報の全て」という意味があったことを思い出しましょう. とくに (中略) 事象 $\{Y = t\}$ はつねに $\mathcal{F}$ に含まれました. したがって, **情報 $\mathcal{F}$ のすべてにアクセス可能なら, そのときには Y の実現値が観測可能ということになります.**』 (強調は前川)

田中 久稔, 計量経済学のための数学, p.183

『$\mathcal{F}$ よりも粗い情報である $\mathcal{G} \subset \mathcal{F}$ のもとでは, Y の実現値を直接観察できるとは限りません. そのときには, **手に入る情報 $\mathcal{G}$ を活用して Y の代用品を見つける**しかありません. 条件付き期待値 $\mathbb{E}[Y \mid \mathcal{G}]$ が, その代用品だというわけです.』(強調は前川)

以上の議論は, 以下のように結論づけられる.

田中 久稔, 計量経済学のための数学, p.183

『定義 9.4 は, 「条件付き期待値 $\mathbb{E}[Y \mid \mathcal{G}]$ は, 限られた情報 $\mathcal{G}$ のもとでの Y の代用品である」 ということを述べています.』

■「合理的期待」　さらに例として, マクロ経済学の合理的期待を例示して **Def 9.4** の意味が説明される.

Assumption: 合理的期待

$$\mathbb{E}\big[\mathbb{E}[Y_{t+1} \mid \mathcal{G}_t]\big] = \mathbb{E}[Y_{t+1}]$$

t 期の principle にとって, 『観測によって明らかになる情報のすべて』 $\mathcal{F}$ は未来の情報も含み, 完全には入手できない. 代わりに利用できるのは t 期までの情報 $\mathcal{G}_t \subset \mathcal{F}$ に限られる. 期待値 $\mathbb{E}[Y_{t+1} \mid \mathcal{G}_t]$ が $\mathcal{G}_t$ 可測であることが **条件 1** だが, これはより細かな情報 $\mathcal{G}_{t+1}$ を principle が利用することを, $\mathcal{G}_t \subset \mathcal{G}_{t+1}$, $\mathcal{G}_{t+1} \not\subset \mathcal{G}_t$ の包含関係から禁止していると読み取れる. 次に **条件 2** についてもその意味を確認しよう. $A = \Omega$ のケースにおいて, $\mathbb{E}[\mathbb{E}[Y_{t+1} \mid \mathcal{G}_t]] = \mathbb{E}[Y_{t+1}]$ が条件の書き換えだが, 将来の変数 Y_{t+1} の期待値を, 現時点 t の情報 $\mathcal{G}_t$

に基づいて見積もり, その見積もりの平均をとったものが左辺にあたるが, これは右辺の将来の変数そのものの期待値と一致する, ということである. これは合理的期待の核心的な前提であり, 「予想は体系的に偏らない」ことを意味する. この点で, **条件 2** は, 「条件付き期待値が“整合的”である」ことを数学的に保証しており, 経済学的には「期待が体系的に誤っていない」という仮定と一致する. 合理的期待の仮定とは, このように確率論的に定義された条件付き期待値の性質を, 経済主体の行動に当てはめることである.

■一般化繰り返し期待値の法則　いよいよ LIE の一般化を示す.

Thm 9.6: Tower Property (塔の性質), 一般化 LIE

$\mathcal{H} \subset \mathcal{G} \subset \mathcal{F}$ をすべて σ-加法族とし, 以下が成り立つ:

$$\mathbb{E}\big[\mathbb{E}[Y \mid \mathcal{G}] \mid \mathcal{H}\big] = \mathbb{E}\big[\mathbb{E}[Y \mid \mathcal{H}] \mid \mathcal{G}\big] = \mathbb{E}[Y \mid \mathcal{H}]$$

Proof $\mathbb{E}[\mathbb{E}[Y \mid \mathcal{G}]\mathcal{H}] = \mathbb{E}[Y \mid \mathcal{H}]$ を証明したい. まず $W = \mathbb{E}[Y \mid \mathcal{G}]$ とおく. **Def 9.4** によって W は $\mathcal{G}$ 可測 ($\because$ **条件 1**) かつ任意の $A \in \mathcal{G}$ について $\mathbb{E}[\mathbb{1}_A W] = \mathbb{E}[\mathbb{1}_A Y]$ を満たす ($\because$ **条件 2**). 次に $V = \mathbb{E}[W \mid \mathcal{H}]$ とおく. 同様に V は $\mathcal{H}$ 可測 ($\because$ **条件 1**) かつ任意の $B \in \mathcal{H}$ について $\mathbb{E}[\mathbb{1}_B V] = \mathbb{E}[\mathbb{1}_B W]$ が成り立つ ($\because$ **条件 2**). ここで仮定 $\mathcal{H} \subset \mathcal{G}$ より任意の $B \in \mathcal{H}$ は $B \in \mathcal{G}$.
以上から:

$$\mathbb{E}[\mathbb{1}_B V] = \mathbb{E}[\mathbb{1}_B W] = \mathbb{E}[\mathbb{1}_B Y].$$

よって $\mathbb{E}[Y \mid \mathcal{H}] = V = \mathbb{E}\big[W \mid \mathcal{H}\big] = \mathbb{E}\big[\mathbb{E}[Y \mid \mathcal{G}] \mid \mathcal{H}\big]$ が得られる.
次に, $\mathbb{E}[\mathbb{E}[Y \mid \mathcal{H}] \mid \mathcal{G}] = \mathbb{E}[Y \mid \mathcal{H}]$ を示したい. $\mathcal{H} \subset \mathcal{G}$ であるから, $Z = \mathbb{E}[Y \mid \mathcal{H}]$ は $\mathcal{G}$ 可測. ここで, $X = Z$ について, X も $\mathcal{G}$ 可測である (**条件 1**). さらに, 任意の $A \in \mathcal{G}$ について $\mathbb{E}[\mathbb{1}_B X] = \mathbb{E}[\mathbb{1}_B Z]$ が成立する (**条件 2**). 以上から X は条件付期待値としての条件を満たし, **Def 9.4** から:

$$\mathbb{E}[Y \mid \mathcal{H}] = Z = X = \mathbb{E}[Z \mid \mathcal{G}] = \mathbb{E}[\mathbb{E}[Y \mid \mathcal{H}] \mid \mathcal{G}]$$

が成立する. よって, 目標の三つの条件付期待値が等式で結ばれる. □

最も粗い情報 $\mathcal{H}$ によって条件付期待値は左右されている.

Thm 9.7: $\mathbb{E}[Y \mid \mathcal{G}]$ の性質

Y, W を確率空間上の確率変数, $\mathcal{G}$ を $\mathcal{F}$ の部分 σ-加法族とするとき, 以下が成立する:

1. 定数 c について $\mathbb{E}[c \mid \mathcal{G}] = c$,
2. 任意の定数 a, b について, $\mathbb{E}[aY + bW \mid \mathcal{G}] = a\mathbb{E}[Y \mid \mathcal{G}] + b\mathbb{E}[W \mid \mathcal{G}]$,
3. W が $\mathcal{G}$-可測であるとき, $\mathbb{E}[YW \mid \mathcal{G}] = W \cdot \mathbb{E}[Y \mid \mathcal{G}]$.

3 の証明には優収束定理が必要らしい. (証明: **問題 9.5**, 未確認)

Def 9.5: 確率変数への条件付期待値

$(\Omega, \mathcal{F}, P)$ 上の確率変数 Y と, 同空間上の確率変数ベクトル $\mathbf{X} : \Omega \to \mathbb{R}^k$ に対し, $\mathcal{G} = \sigma[\mathbf{X}] \subset \mathcal{F}$ 上に条件付けられた, 条件付期待値 $\mathbb{E}[Y \mid \mathcal{G}]$ を Y の $\mathbf{X}$ に条件付けられた期待値といい, $\mathbb{E}[Y \mid \mathbf{X}]$ と書く.

未確認

Thm 9.8: 条件付期待値の性質

確率空間 $(\Omega, \mathcal{F}, \mathbf{P})$ 上の確率変数 Y, W と確率変数ベクトル $\mathbf{X}$ に対し, $\mathbb{E}[\,Y \mid \mathbf{X}\,]$ は以下を満たす:

1. 定数 c について $\mathbb{E}[\,c \mid \mathbf{X}\,] = c$
2. 任意の定数 a, b について $\mathbb{E}[\,aY + bZ \mid \mathbf{X}\,] = a\,\mathbb{E}[\,Y \mid \mathbf{X}\,] + b\,\mathbb{E}[\,Z \mid \mathbf{X}\,]$
3. W が $\mathbf{X}$ 可測のとき $\mathbb{E}[\,W \cdot Y \mid \mathbf{X}] = W \cdot \mathbb{E}[\,Y \mid \mathbf{X}\,]$
4. $\mathbb{E}\big[\,\mathbb{E}[\,Y \mid W, \mathbf{X}] \mid \mathbf{X}\,\big] = \mathbb{E}[\,Y \mid \mathbf{X}\,]$
5. $\mathbb{E}\big[\mathbb{E}[Y \mid \mathbf{X}]\big] = \mathbb{E}[Y]$

上の諸性質は, **Thm 9.7** に確率変数から生成された σ-加法族を対応させたものとして得られる. とくに 3 については, 適当な関数 g を用いて $W = g(\mathbf{X})$ と表現されるものは $\mathbf{X}$ 可測なので, **Thm 9.3** に対応する結果となる $\mathbb{E}[\,g(\mathbf{X}) \cdot Y \mid \mathbf{X}] = g(\mathbf{X}) \cdot \mathbb{E}[\,Y \mid \mathbf{X}\,]$ が得られる.

C.4 大数の法則と推定量の一致性 (10 章)

■独立性 『計量経済学のための数学』の 10 章 1 節を参考にして独立性を定義しておく.

Def 10.1: 事象の独立性

確率空間 $(\Omega, \mathcal{F}, \mathbf{P})$ 上の事象 $A_1, \ldots, A_n \in \mathcal{F}$ が互いに独立であるとは, 以下が成り立つことをいう:

$$\mathbf{P}\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} \mathbf{P}A_i$$

Def 10.2.1: 集合族の独立性

確率空間 $(\Omega, \mathcal{F}, \mathbf{P})$ 上の部分集合族 $\mathcal{G}_1, \ldots, \mathcal{G}_n \subset \mathcal{F}$ が互いに独立であるとは, これらの集合族から選ばれた任意の事象 $A_1 \in \mathcal{G}_1, \ldots, A_n \in \mathcal{G}_n$ がつねに独立になることをいう.

Def 10.2.2: 確率変数ベクトルの独立性

確率空間 $(\Omega, \mathcal{F}, \mathbf{P})$ 上の確率変数 $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ が独立であるとは, これらが生成する σ-加法族 $\sigma[\mathbf{Z}_i] \subset \mathcal{F}$ に関して, $\sigma[\mathbf{Z}_1], \ldots, \sigma[\mathbf{Z}_n]$ が独立であることを指す. とくに, $\mathbf{Z}_1, \mathbf{Z}_2$ が独立なことを以下のように表す:

$$\mathbf{Z}_1 \perp\!\!\!\perp \mathbf{Z}_2$$

Thm 10.1: 確率変数ベクトルの独立性と確率測度

$\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ が独立である必要十分条件として, 任意の $\mathbf{t}_1, \ldots, \mathbf{t}_n$ で以下が成立することがある:

$$\mathbf{P}\{\mathbf{Z}_1 \le \mathbf{t}_1, \ldots, \mathbf{Z}_n \le \mathbf{t}_n\} = \mathbf{P}\{\mathbf{Z}_1 \le \mathbf{t}_1\} \cdots \mathbf{P}\{\mathbf{Z}_n \le \mathbf{t}_n\}$$

Corr 10.1: 連続確率変数における独立性

連続確率変数 $X_1, \ldots, X_n$ と, これを成分とするベクトル $\mathbf{X}$ が, 周辺密度 $f_{X_i}(t_1)$ と, 結合密度関数 $f_{\mathbf{X}}(\mathbf{t})$ を持つとする. このとき, 次が成り立つことが確率変数の独立性の必要十分条件である:

$$f_{\mathbf{X}}(\mathbf{t}) = \prod_{i=1}^{n} f_{X_i}(t_i)$$

Def: 独立同一な確率収束列, i.i.d. 性

確率変数列 $\{\mathbf{Z}_n\}_{n=1}^{\infty}$ が以下を満たすとき, これは独立同一に分布する (i.i.d.) 確率変数列という:

1. 任意の $n \in \mathbb{N}$ に対して, $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ は独立である.
2. 同一の分布関数 $F : \mathbb{R}^k \to [0,1]$ が存在して, 全ての $\mathbf{t} \in \mathbb{R}^k$ について, 以下が成立する:

$$\mathbf{P}\{\mathbf{Z}_1 \leq \mathbf{t}\} = \mathbf{P}\{\mathbf{Z}_2 \leq \mathbf{t}\} = \cdots = F(\mathbf{t})$$

Def 10.3: 確率収束 (ベクトル列)

確率変数ベクトル列 $\{\mathbf{W}_n\}$ が $\mathbf{a} \in \mathbb{R}^k$ に確率収束するとは, 任意の $\varepsilon > 0$ に対して

$$\lim_{n\to\infty} \mathbf{P}\big\{\|\mathbf{W}_n - \mathbf{a}\| < \varepsilon\big\} = 1$$

が成立することであり, 確率収束を $\mathbf{W}_n \xrightarrow{p} \mathbf{a}$ と表記する.

Thm 10.3: 連続写像定理 (CMT)

確率変数ベクトル列 $\mathbf{W}_n \xrightarrow{p} \mathbf{a}$ と $\mathbf{a} \in \mathbb{R}^d$ において連続な関数 $f : \mathbb{R}^k \to \mathbb{R}$ について, 以下が成り立つ:

$$f(\mathbf{W}_n) \xrightarrow{p} f(\mathbf{a})$$

関数の連続性, 確率測度の単調性, 確率収束の定義により示される.

Def: 確率収束 (行列)

確率行列列 $\{\mathbf{A}_n\} \subset \mathbb{R}^{k\times l}$ が行列 $\mathbf{A}$ に確率収束するとは, $\mathbf{A}_n$ の各成分 $a_{n,ij}$ がそれぞれ $\mathbf{A}$ の各成分 a_{ij} に確率収束することを指し, 以下のように表記する:

$$\mathbf{A}_n \xrightarrow{p} \mathbf{A}$$

Thm: 逆行列の確率収束

確率行列列 $\mathbf{A}_n \xrightarrow{p} \mathbf{A}$ かつ $\det(\mathbf{A}) \neq 0$ のとき,

$$\mathbf{A}_n^{-1} \xrightarrow{p} \mathbf{A}^{-1}$$

CMT によって示される. ただし, 確率変数ベクトルの確率収束とその各成分の確率収束が必要十分条件の関係にあることを利用する必要がある筈, 教科書では特に記載がない.

Lemma 10.1: チェビシェフの不等式

任意の確率変数 X と, 任意の $\varepsilon > 0$ に関して, 以下が成り立つ:

$$\mathbf{P}\big\{|X| \geq \varepsilon\big\} \leq \frac{\mathbb{E}[X^2]}{\varepsilon^2}$$

チェビシェフの不等式によって以下の定理が示される.

Thm 10.5: 大数の法則 (WLLN, LLN)

i.i.d. な確率変数列 $\{Z_i\}_{i=1}^n$, $Var(Z) < \infty$ について, 標本平均 $\bar{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i$ は以下を満たす:

$$\bar{Z}_n \xrightarrow{p} \mathbb{E}[Z]$$

Thm 10.6: 大数の法則 (ベクトル列)

i.i.d な確率変数ベクトル列 $\{\mathbf{Z}_i\} \subset \mathbb{R}^k$ について $\mathbb{E}|Z_1|, \ldots, \mathbb{E}|Z_k| < \infty$ であるとき, $\overline{\mathbf{Z}}_n$ は以下を満たす:

$$\overline{\mathbf{Z}}_n \xrightarrow{p} \mathbb{E}[\mathbf{Z}]$$

標本平均も確率変数である. 一般の推定量も定義より確率変数である.

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
[5] Hansen, B.E. (2022), *Econometrics*, Princeton University Press
[6] InsightEdge, データ駆動型回帰分析を実装してみた, `https://techblog.insightedge.jp/entry/non-semipara`, 2025/06/18 取得