令和 7 年度　中級計量経済学
講義ノート 2: 線形回帰モデル

このノートでは、線形回帰モデルの理論を紹介する。回帰モデルは、ある変数から別の変数への影響を調べたいときに用いられる。こうした変数間の関係を線形に表現して、確率論的な仮定を置くことで統計学的な分析を可能にしたものが線形回帰モデルである。線形回帰モデルは、経済分析で最もよく使われるモデルであると同時に、他の分析手法の基礎となっているため、計量経済学において最も重要なトピックである。実際の計量経済分析では、最小二乗法を用いてこのモデルを推定し、その結果を用いて検定や統計的推測が行われる。この章では、そうした統計手法の理論的背景を学習することである。加えて、近年進歩が著しい因果効果の分析手法についても触れる。

2.1　線形回帰モデルと最小二乗法

$(y_i, \mathbf{X}_i)$ という変数の組を $i = 1, \dots, n$ について観測したとしよう。変数 y_i と $\mathbf{X}_i$ との関係が線形であると想定し、

$$y_i = \beta' \mathbf{X}_i + u_i \tag{1}$$

というモデルを考える。u_i は誤差項と呼ばれ、y_i を決める要素のうち $\mathbf{X}_i$ 以外のものをまとめたものである。上のモデルは、線形モデルと呼ばれる。線形回帰モデルと呼ぶ際には、通常、暗黙に $\mathbf{X}_i$ と u_i が無相関であるという仮定、または $\mathbf{X}_i$ を条件とする u_i の期待値が 0 であるという追加的な仮定をおく。これらについては、後に詳しく説明する。

変数の組のベクトル $\mathbf{X}_i$ の最初の要素は通常 1 とする、つまり $\mathbf{X}_i = (1, x_{i2}, \dots, x_{ik})'$ である。係数のパラメータ（母数）ベクトルを $\beta' = (\beta_1, \beta_2, \dots, \beta_k)$ とすると、

$$\beta' \mathbf{X}_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \tag{2}$$

となり、最初の要素 1 は定数項に対応する。

用語　以下の用語は、講義を通して繰り返し使用する。

- β_1: 定数項、切片と呼ばれる。
- $(\beta_2, \dots \beta_k)$: 係数と呼ばれる。
- $y = \beta' \mathbf{X}$: 母回帰線。
- β: 母回帰線の係数（母数）。
- y_i: 従属変数、被説明変数。
- $\mathbf{X}_i$: 独立変数、説明変数、共変量、回帰変数など、分野によっていろいろな呼び方があり、どの呼び方も使用されている。各名称ごとに少しずつニュアンスは異なるが、その違いはそれほど気にすることはないと思われる。
- u_i: 誤差項と呼ばれる。しかし、経済分析においては、文字通りの誤差（測定や観測の誤差）であるという理解は必ずしも適切でない。y_i を決める決定要因のうち $\mathbf{X}_i$ 以外のものすべてを含んだ量と解釈するべきである。y_i への影響度の点では $\mathbf{X}_i$ よりも重要なものを含む可能性もある。このようなことが起こる理由は、データが入手できない、そもそも重要な決定要因を見落としているといったことが考えられる。

線形回帰モデルの係数の推定　回帰モデルの分析では、未知母数のベクトル β が最大の興味対象であり、これをデータから推定する方法とその性質について説明する。

$\mathbf{X}_i$ を使って y_i を予測するときの誤差の 2 乗和を最小化するやり方を最小二乗推定という。また、その推定量を、最小二乗推定量という。これを OLS(ordinary least squares) と表記する。β の OLS 推定量を $\hat{\beta}$ とすると、それは、

$$\sum_{i=1}^{n}(y_i - b'\mathbf{X}_i)^2 \tag{3}$$

を最小化する b の値である。1 次の条件は、

$$-\sum_{i=1}^{n}\mathbf{X}_i(y_i - b'\mathbf{X}_i) = 0 \tag{4}$$

となる。目的関数は凸関数であるので、1 次の条件は最小化のための必要十分条件になっている。

OLS 推定量 $\hat{\beta}$ は、

$$\hat{\beta} = \left(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\sum_{i=1}^{n}\mathbf{X}_i y_i \tag{5}$$

と明示的に書くことができる。

以下の用語を講義を通して使用する。

- $y = \hat{\beta}'\mathbf{X}$: OLS 回帰直線。
- $\hat{y}_i = \hat{\beta}'\mathbf{X}_i$: OLS 回帰線による $X = X_i$ の時の y_i の予測値。
- $\hat{u}_i = y_i - \hat{y}_i$: i 番目の観測値の残差。これは、誤差項 (u_i) とは違うことに注意しよう。

2.2　線形回帰モデルの仮定と OLS 推定量の統計的性質

推定量の統計的な性質を明らかにすることが、この章で最も重要なポイントである。そのために以下の仮定をおく。その仮定の下で、不偏性、一致性、漸近正規性という 3 つの性質を解説する。

最小二乗法の性質を調べるための仮定

1. $E(u_i|\mathbf{X}_i) = 0$ あるいは $E(y_i|X_i) = \beta'\mathbf{X}_i$。ここで、$E(u_i|\mathbf{X}_i) = 0$ であるとき $corr(\mathbf{X}_i, u_i) = 0$ となる。(逆は必ずしも成り立たない。)
2. $(\mathbf{X}_i, y_i), i = 1, \ldots, n$ は i.i.d. (独立同一分布)。もし、ある母集団から無作為抽出によって観測値を集めた場合、この仮定は満たされる。
3. $\mathbf{X}_i$ と u_i は 4 次のモーメントを持つ。つまり、すべての m について $0 < E((x_{im} - E(x_{im}))^4) < \infty$ となり、また $E(u_i^4) < \infty$ である。これは異常に大きい X_i や u_i の値をそれほど頻繁には取らないことを意味する。
4. $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)'$ という行列は、列フルランクである。完全な多重共線性がない（要素間に確定的な線形関係（例えば $X_{i1} + X_{i2} = X_{i3}$ 等）がない）ことを仮定している。

これらの仮定の役割

- 数学的に、OLS 推定量が一致性、不偏性、漸近正規性をもつことを示すのに使われる。
- OLS による回帰分析がうまく機能しない状況を明らかにしたり、その対応策を考える際に役に立つ。最初の仮定 ($E(u_i|\mathbf{X}_i) = 0$) が、実証研究においては、もっともよく議論される（次章)。

OLS 推定の漸近的性質　OLS 推定量 $\hat{\beta}$ は、ランダムに抽出された標本から計算している。したがって、推定量は、ある確率分布を持つ確率変数である。

OLS 推定量の分布を知ることは、母数の値に関する仮説検定や、信頼区間の構築に必要となる。

誤差項 u_i の分布を決めてしまわずに議論するために、大標本理論 (n を無限に大きくした極限で考える統計理論) を使って推定量の標本分布を近似する。

OLS 推定量は、

- 不偏 ($E(\hat{\beta}) = \beta$) で、
- 一致 ($\hat{\beta} \to_p \beta$) で、
- 漸近正規:

$$\sqrt{n}(\hat{\beta} - \beta) \to_d N\left(\mathbf{0}, \left(E(\mathbf{X}_i\mathbf{X}_i')\right)^{-1} E(u_i^2\mathbf{X}_i\mathbf{X}_i') \left(E(\mathbf{X}_i\mathbf{X}_i')\right)^{-1}\right). \tag{6}$$

である。

証明

Proof. **不偏性:** $\hat{\beta}$ の式 (5) に $y_i = \beta'\mathbf{X}_i + u_i$ を代入すると

$$\hat{\beta} = \beta + \left(\sum_{i=1}^{n} \mathbf{X}_i\mathbf{X}_i'\right)^{-1} \sum_{i=1}^{n} \mathbf{X}_i u_i \tag{7}$$

が成り立つことがわかる。ここで、i.i.d. の仮定と、$E(u_i|\mathbf{X}_i) = 0$ の仮定から、

$$E\left\{\left(\sum_{i=1}^{n} \mathbf{X}_i\mathbf{X}_i'\right)^{-1} \sum_{i=1}^{n} \mathbf{X}_i u_i|\mathbf{X}\right\} = \left(\sum_{i=1}^{n} \mathbf{X}_i\mathbf{X}_i'\right)^{-1} \sum_{i=1}^{n} \mathbf{X}_i E(u_i|\mathbf{X}_i) = 0 \tag{8}$$

となる。従って、繰り返し期待値の法則（数学付録 p.4 参照）から、$E(\hat{\beta}) = E\{E(\hat{\beta}|X)\} = E(\beta) = \beta$ となる。

一致性: まず、

$$\hat{\beta} = \beta + \left(\sum_{i=1}^{n} \mathbf{X}_i\mathbf{X}_i'\right)^{-1} \sum_{i=1}^{n} \mathbf{X}_i u_i = \beta + \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i\mathbf{X}_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i u_i \tag{9}$$

と表現できる。大数の法則（数学付録 p.6 参照）より、

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i\mathbf{X}_i' \to_p E(\mathbf{X}_i\mathbf{X}_i') \tag{10}$$

となり、右辺は正値定符号である。また、大数の法則と繰り返し期待値の法則から

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i u_i \to_p E(\mathbf{X}_i u_i) = E\{\mathbf{X}_i E(u_i|\mathbf{X}_i)\} = 0 \tag{11}$$

も示せる。つまり、$\hat{\beta} \to_p \beta$ となる。ここまでの証明で、$E(u_i|\mathbf{X}_i) = 0$ の仮定が重要な意味をもっていることがわかるであろう。これが満たされない場合には、不偏性や一致性が成立しなくなってしまう。。

漸近正規性: まず、(7) 式を変形すると

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_i u_i. \tag{12}$$

と表すことができる。中心極限定理（数学付録 p.6 参照）により、

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_i u_i \to_d N\left(\mathbf{0}, E(u_i^2\mathbf{X}_i\mathbf{X}_i')\right) \tag{13}$$

となる。したがって、Slutsky の補題（数学付録 p.6 参照）と正規分布の性質（数学補論 p.4 参照）により

$$\begin{aligned}
\sqrt{n}(\hat{\beta} - \beta) &\to_d \left(E(\mathbf{X}_i\mathbf{X}_i')\right)^{-1} N\left(\mathbf{0}, E(u_i^2\mathbf{X}_i\mathbf{X}_i')\right) && (14)\\
&= N\left(\mathbf{0}, \left(E(\mathbf{X}_i\mathbf{X}_i')\right)^{-1} E(u_i^2\mathbf{X}_i\mathbf{X}_i')\left(E(\mathbf{X}_i\mathbf{X}_i')\right)^{-1}\right). && (15)
\end{aligned}$$

□

漸近分散の推定　OLS 推定量の漸近分散は、次のように推定できる。

$$\hat{\mathbf{V}} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2\mathbf{X}_i\mathbf{X}_i'\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}. \tag{16}$$

この推定量は、上で述べた仮定の下で一致性 ($\hat{\mathbf{V}} \to_p \left(E(\mathbf{X}_i\mathbf{X}_i')\right)^{-1} E(u_i^2\mathbf{X}_i\mathbf{X}_i')\left(E(\mathbf{X}_i\mathbf{X}_i')\right)^{-1}$) を持つ。証明は省略する。

分散均一の場合　もし $E(u_i^2|\mathbf{X}_i) = \sigma^2$ となるなら、分散均一であるという。そのとき、上の結果はすべて成立するが、漸近分散の表現が簡単になり、

$$\sqrt{n}(\hat{\beta} - \beta) \to_d N\left(\mathbf{0}, \sigma^2\left(E(\mathbf{X}_i\mathbf{X}_i')\right)^{-1}\right) \tag{17}$$

となる。また、漸近分散の推定も $s^2 = \sum_{i=1}^{n}\hat{u}_i^2/n$ として、

$$\hat{\mathbf{V}} = s^2\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1} \tag{18}$$

となる。ここに示した分散均一の場合の漸近分散の方が (16) 式の一般形よりも簡潔なので、多くの統計学、計量経済学のテキストの回帰分析の章では、こちらが紹介されている。

2.3 仮説検定と信頼区間

ある一つの回帰係数に関する検定　m 番目の回帰変数の係数 β_m に関する帰無仮説 $H_0 : \beta_m = \beta_{m,0}$ は、t 統計量を使って検定できる。ここで、$\beta_{m,0}$ は経済理論等から導かれる特定の値である。まず帰無仮説が正しいにも関わらず間違ってこれを棄却する確率（これを有意水準という）を決める。通常、有意水準は５％、１％などの小さい値とする。

1. OLS 推定値 $\hat{\beta}_m$ を計算する。
2. $\hat{\beta}_m$ の標準誤差を計算する。$\hat{\mathbf{V}}_{mm}$ を、$\hat{\mathbf{V}}$ の (m, m) 要素とする。$\hat{\beta}_m$ の標準誤差 $SE(\hat{\beta}_m)$ は
$$SE(\hat{\beta}_m) = \sqrt{\frac{\hat{\mathbf{V}}_{mm}}{n}} \tag{19}$$
である。
3. t 統計量を計算する。
$$t = \frac{\hat{\beta}_m - \beta_{m,0}}{SE(\hat{\beta}_m)}. \tag{20}$$
4. 有意水準を、たとえば5%と決め、もし、$|t| > 1.96$ ならば、H_0 を棄却する。あるいは、p 値を $2\Phi(-|t|)$ として計算して、これが 5% より小さければ H_0 を棄却する。。

この手続きの意味を理解するために、t 統計量を以下のように書き換えてみよう。

$$t = \frac{\hat{\beta}_m - \beta_{m,0}}{\mathrm{SE}(\hat{\beta}_m)} = \frac{\hat{\beta}_m - \beta_m}{\mathrm{SE}(\hat{\beta}_m)} + \frac{\beta_m - \beta_{m,0}}{\mathrm{SE}(\hat{\beta}_m)} \tag{21}$$

最小二乗推定量の漸近正規性から、右辺の第１項は常に近似的に標準正規分布に従うので、95％の確率でおよそ-2 と 2 の間の値をとる。一方、第 2 項は帰無仮説が正しい時には 0 であるが、対立仮説が正しければ 0 でない。(19) 式から、n が大きくなれば分母の $\mathrm{SE}(\hat{\beta}_m)$ は 0 に近づいていくため、第 2 項の絶対値はどんどん大きくなる。まとめると、帰無仮説が正しい時にはおよそ確率 0.95 で $|t| < 1.96$ となり、対立仮説が正しい時には $|t|$ はそれよりもずっと大きな値をとるはずである。

片側検定　帰無仮説を $H_0 : \beta_m = \beta_{m,0}$ とし、対立仮説を $H_1 : \beta_m < \beta_{m,0}$ とする。両側検定と片側検定の違いは、手順の 4 番にある。

- p 値: $\Phi(t)$.
- 有意水準が 5%なら $t < -1.645$ のとき帰無仮説を棄却する。

回帰係数の信頼区間　β_m の 95%信頼区間とは、

- 両側検定をしたときに、5%の有意水準では棄却できない帰無仮説のもとでの係数の値の集合。
- 95%の確率で、β_m の真の値を含む区間。ここで、区間が確率変数である。

95%信頼区間は、近似的に

$$P(|\frac{\hat{\beta}_m - \beta_m}{\mathrm{SE}(\hat{\beta}_m)}| < 1.96) = 0.95 \tag{22}$$

が成り立つことから、不等式を変形して

$$(\hat{\beta}_m - 1.96 \times SE(\hat{\beta}_m), \quad \hat{\beta}_m + 1.96 \times SE(\hat{\beta}_m)) \tag{23}$$

となる。

x を変化させたときの y の変化分の予測値の信頼区間　x_m を Δx_m だけ変化させると、y の変化分の予測値は、$\beta_m \Delta x_m$ である。

つまり、$\beta_m \Delta x_m$ の信頼区間が必要となる。これは

$$(\hat{\beta}_m \Delta x_m - 1.96 \times SE(\hat{\beta}_m)|\Delta x_m|, \quad \hat{\beta}_m \Delta x_m + 1.96 \times SE(\hat{\beta}_m)|\Delta x_m|) \tag{24}$$

として計算できる。

2.4 複合仮説の検定

複合仮説とは、二つ以上の制約のある仮説である。主に、二つ以上の係数がそれぞれある特定の値であるという仮説を考える。

Wald 統計量と F 統計量　帰無仮説を、$R\beta - r = 0$ とする。ここで、R は $q \times k$ の行列で、行フルランクであり、r は $q \times 1$ のベクトルとする $(q < k)$。

例えば、$\beta = (\beta_1, \ldots, \beta_4)$ であり、帰無仮説が $H_0 : \beta_1 = \beta_2, \beta_3 = 0$ であるとすると、それに対応する R と r は、

$$R = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad r = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{25}$$

である。

Wald 統計量は、

$$W = n(R\hat{\beta} - r)' \left(R\hat{V}R' \right)^{-1} (R\hat{\beta} - r) \tag{26}$$

である。H_0 のもとで、$W \to_d \chi^2_q$ である。

F 統計量は、$F = W/q$ である。

- 注：分散不均一に頑健な Wald あるいは F 統計量を使うこと。多くの統計ソフトでは、特に指定しない限り、分散均一の場合のみ使用できる Wald あるいは F 統計量が計算される。

- q は、帰無仮説を成り立たせる最小の制約の数である。例えば、$\beta_1 = \beta_2 = \beta_3$ という帰無仮説の場合は、制約の数は、$q = 2$ となる。

- 例えば、$(\beta_1, \beta_2) = (0, 0)$ といった二つのパラメータの値について検定を行いたいとする。そのとき、それぞれのパラメータごとに t 検定を行って、どちらかが棄却されたら $(\beta_1, \beta_2) = (0, 0)$ という帰無仮説を棄却するというやり方は理にかなっているように思われる。しかし、実はこのアプローチでは検定全体の有意水準が意図した値とはずれてしまうという問題が生ずる。今、それぞれについて有意水準を 5 %とする検定を行うとする。t_1 か t_2 のどちらかの絶対値が、1.96 を超えたときに複合仮説を棄却すると、全体の有意水準はどうなるであろうか。

  仮に t_1 と t_2 が独立である場合には、どちらも棄却されない確率は
$$\Pr(|t_1| < 1.96, |t_2| < 1.96) = \Pr(|t_1| < 1.96) \times \Pr(|t_2| < 1.96) = 0.95^2 = 0.9025 \tag{27}$$
  となり、帰無仮説が正しいにも関わらず棄却される確率は、$\Pr(\text{棄却} |H_0) = 1 - 0.9025 = 0.0975$ となる。これは 5 %の 2 倍に近く、帰無仮説は必要以上に高い確率で棄却されることになる。一方で、t_1 と t_2 が完全に相関しているなら、有意水準は 5% になることが確かめられる。つまり、検定全体の有意水準は、t_1 と t_2 の相関に依存してしまい、一般に 5 %にならない。逆に言うと、全体の有意水準を 5% にするためには、t_1 と t_2 の相関に応じて個別の t 検定における棄却域をうまく設定しなければならない。

- β_1 を切片として、次の帰無仮説の F 統計量を考える。
$$H_0 : \beta_2 = 0, \ldots, \beta_k = 0. \tag{28}$$
  この場合、$R = [\mathbf{0}_{(k-1)\times 1} : I_{k-1}]$ で $r = \mathbf{0}_{(k-1)\times 1}$ である。この統計量は、回帰全体の F 統計量と呼ばれ、多くの統計ソフトで回帰をすると自動的に計算される。

複数の係数に関する信頼集合　二つ以上の係数に関する 95%の信頼集合とは、95%の確率で真の係数値を含む集合である。

- これは、5%の有意水準で、係数の組がある値であるという F 検定が棄却できない、係数の値の集合である。

- 2 つの係数の時には、信頼集合は楕円になる。

- 近年では、多くの統計ソフトで計算できるようになってきた。

2.5　当てはまりの良さ：　R^2

決定係数 R^2 とは、y_i の標本分散のうち、$\mathbf{X}_i$ で説明できる割合である。

まず、次のような y_i の分割を考える: $y_i = \hat{y}_i + \hat{u}_i$。

- $\hat{y}_i = \hat{\beta}'\mathbf{X}_i$: y_i のうち、モデルによって説明できる部分。

- $\hat{u}_i = y_i - \hat{y}_i$: y_i のうち、モデルで説明できない部分。

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n \hat{u}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}. \tag{29}$$

$0 \le R^2 \le 1$ である。R^2 は、回帰変数を増やすと増加する。

R^2 が 1 に近いということは、回帰変数が Y_i の値を予測する精度が高いということである。一方で、R^2 が 0 に近いということは、回帰変数が Y_i を予測するのにあまり役に立たないということである。

修正済み R^2　説明変数の追加によって R^2 が大きくなったからといって、その変数がモデルの当てはまりを改善したとは一概には言えない。実は、R^2 は回帰変数を増やすと常に大きくなる（もしくは変化しない）。この問題に対処するために、修正済み R^2 を使う。

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{\sum_{i=1}^n \hat{u}_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}. \tag{30}$$

回帰変数を増やすと $\sum_{i=1}^n \hat{u}_i^2$ は減るが、$1/(n-k-1)$ は大きくなる。

- $\bar{R}^2 \leq R^2$.
- $\bar{R}^2$ は負になることもある。

注意すべき点

- R^2 が増えたからといって、追加した変数が統計的に有意とは限らない。
- R^2 や $\bar{R}^2$ が大きいからと言って、回帰変数が、被説明変数を決める真の要因になっているとはいえない。

2.6　分散均一性と不均一性

- 分散均一性: u_i の条件付き分散が $\mathbf{X}_i$ に依存していないということ。($E(u_i^2|\mathbf{X}_i) = \sigma^2$)
- 分散不均一性: u_i の条件付き分散が $\mathbf{X}_i$ に依存しているということ。

例として、次のような回帰モデルを考える。$Earnings$ は所得であり、$MALE$ は男性なら 1、女性なら 0 をとる 2 項変数とする。

$$Earnings_i = \beta_1 + \beta_2 MALE_i + u_i. \tag{31}$$

分散均一性とは、所得の分散が男性と女性で同じであるということ。

分散均一性の仮定の功罪

- 制約が強い。
- OLS は、分散均一性が成り立っていれば、有効推定量（最小分散）になる。“The Gauss-Markov theorem”。
- OLS 係数推定値の標準誤差を簡単な式で計算できる ((18) 式から)。

統計的推測（検定や区間推定）には必ず標準誤差が用いられる。(16) 式から計算した標準誤差は、分散不均一に対して頑健なものである。つまり、分散不均一であってもなくても漸近的に正しい値を与える。したがって、その標準誤差を使用することで、誤差項の分散均一の是非に関わらず、適切に統計的推測ができる。逆に、(18) 式は分散均一の時は正しいが、分散不均一の時は間違った結果を与える。

どちらを使うべきか?　自然科学の実験と違って、経済分析では分散均一でない場合が多い。したがって、分散不均一に頑健な標準誤差を使うのがよいであろう。

なお、多くの統計ソフトの回帰分析では、デフォルトでは分散均一の場合のみ使える標準誤差を計算するので注意すること。

2.7 どの変数を回帰に含めるべきか

経済の実証分析においては、二つの変数の関係、特にある特定の変数が別の変数に与える影響に興味があることが多い。しかし、実際に回帰を行うときは、興味のある変数以外の変数もモデルに含めて推定することが多く、実は影響を正確に測るためにはそれが必要である。その重要な理由は、欠落変数のバイアスを回避するためである。

例えば、小学校において学級の大きさが学力テストの点数に与える影響を調べたい時に、次のような線形回帰モデルを使ったとする。

$$TestScore = \beta_1 + \beta_2 STR + u \tag{32}$$

として、$TestScore$ が学力テストの点数で、STR が教師一人当たりの児童数であるとする。誤差項の u は教師一人当たりの児童数以外に学力テストの成績に影響を与えるよう要素すべてを含んだものである。したがって、u には、以下のようなものが含まれると考えられる。

- 配置されている教師の資質
- 児童の家庭環境
- 最新の教育設備を導入しているか

これらの要素と学級の大きさは、相関していることが多い。この相関が、問題のある結果を出すこともある。

例えば、ある学校は学級の大きさも小さく、学力テストの成績もよかったとする。しかし、その学校に教師の数が多いことから、教育熱心な家庭が校区に居住するようになっているかもしれず、その時には、成績のよさが、学級の大きさから来るのか、それとも家庭環境からくるのかよくわからなくなってしまう。

欠落変数の定義　以下の二つの条件を満たす変数を欠落変数という。

1. 被説明変数に影響を与えているが、モデルの説明変数に含まれていない。
2. モデルに含まれている回帰変数と相関がある。

もし欠落変数があると、OLS 推定量はバイアスをもち、このバイアスを欠落変数バイアスという。１つ目の条件は当然として、２つ目の条件を課す理由が分かりにくいかもしれない。実は、以下に示すように、２つ目の条件が満たされない（相関がない）時にはバイアスが生じないのである。

欠落変数がある場合に推定結果にどのような影響があるか考えてみよう。Y の決定要因は X, Z であるが、Z を含めなかったとしよう。つまり、

$$Y = X\beta + Z\gamma + u,\ E(u|X, Z) = 0 \tag{33}$$

であるが、X のみを説明変数として β を推定すると、

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta + Z\gamma + u) \\ &= \beta + (X'X)^{-1}X'Z\gamma + (X'X)^{-1}X'u\end{aligned}$$

なので、X と Z が相関をもつことから第二項がバイアスとして残ってしまう。

同じことであるが、次のように説明することもできる。欠落変数がある場合には u_i の中に欠落変数 Z_i が含まれることになってしまうため、

$$E(u_i|\mathbf{X}_i) \neq 0 \tag{34}$$

となり、u_i と X_i に相関が生ずる。つまり、仮定 1 が満たされておらず、そのために OLS 推定量が一致性を失ってしまう。

説明変数が一つの場合を考える。仮に u_i と X_i が相関しているとする。そのとき、OLS 推定量 $\hat{\beta}_1$ の極限は

$$\hat{\beta}_1 \rightarrow_p \beta_1 + \rho_{Xu}\frac{\sigma_u}{\sigma_X} \tag{35}$$

である。ここで、$\rho_{Xu} = corr(X_i, u_i)$ である。

- 欠落変数のバイアスは標本数を増やしても解決しない。
- $|\rho_{Xu}|$ が大きいなら、バイアスも大きい。
- もし ρ_{Xu} が正なら、上向きにバイアスがかかる。

当然のことながら、回帰式に必要な変数（欠落変数）を全て含めれば欠落変数のバイアスを回避できる。しかし、欠落変数に関するデータが入手できない等、それが不可能な場合もある。その場合でも、欠落変数の代わりに次の条件を満たす変数を用意できれば興味のある回帰係数に関する欠落変数バイアスを回避することができる。

条件付き平均に関する独立性　X を興味のある回帰変数、$W_1, W_2, \ldots, W_k$ をバイアス回避のために用いる追加の回帰変数としよう。

興味のあるのは X から y への影響であって、$W_1, \ldots, W_k$ の影響はそれほど興味もない時には、もっと弱い条件の下で X の係数を一致推定できる。それが次にあげる条件付き平均に関する独立性である。$v = y - \beta_0 - \beta_1 X - \alpha_1 W_1 - ... - \alpha_k W_k$ として

$$E(v|X, W_1 \ldots, W_k) = \gamma_0 + \gamma_1 W_1 + \cdots + \gamma_k W_k. \tag{36}$$

ここで重要なのは、この条件付き平均が X に依存していないことである。

なぜこの条件で十分なのかをみるために、X、W、欠落変数 Z がそれぞれ一つだけの簡単な場合を考える。つまり、真の関係が

$$y = \beta_0 + \beta_1 X + \alpha W + \delta Z + u \tag{37}$$

であり ($\alpha = 0$ でも可)、誤差項 u は仮定 1 つまり $E(u|X, W, Z) = 0$ を満たすが、Z が欠落変数であるために、以下のモデルを適用する状況である。

$$y = \beta_0 + \beta_1 X + \alpha W + v \tag{38}$$

新たな"誤差項"$v \equiv \delta Z + u$ は欠落変数を含めたものになっている。ここで、条件付き平均に関する独立性

$$E(v|X, W) = \gamma_0 + \gamma_1 W. \tag{39}$$

が成立しているとすると、$\epsilon \equiv u - \gamma_0 - \gamma_1 W$ とおけば、

$$
\begin{aligned}
y &= \beta_0 + \beta_1 X + \alpha W + v && (40)\\
&= \beta_0 + \beta_1 X + \alpha W + (\gamma_0 + \gamma_1 W + \epsilon) && (41)\\
&= (\beta_0 + \gamma_0) + \beta_1 X + (\alpha + \gamma_1) W + \epsilon && (42)
\end{aligned}
$$

である。u は仮定 1 を満たすので $E(u|X,W)=0$ で（繰り返し期待値の法則）、最後の式に着目すると、誤差項 ϵ は説明変数 X,W を条件として期待値が 0 なので、仮定 1 が満たされていることがわかる。そのため、β_1 は一致推定できるのである。ただし、元の真の関係を表すパラメータ α と β_0 を一致推定することはできない。

2.8 回帰分析の解釈

この節では説明を容易にするために、切片の係数を β_0 とし、最初の回帰変数の係数を β_1 とする。

回帰変数が 2 項変数の場合　2 項変数とは、0 か 1 の二つの値しかとらない変数のことで、ダミー変数とも呼ばれる。

基本的に、回帰変数が 2 項変数でも、そうでないときと全く同様に回帰分析を行うことができる。

しかし、β_1 の解釈は連続変数の場合とは異なる。実は回帰変数が 2 項変数の場合の回帰は、平均の差の分析と同じであることが示される。

D_i を 2 項変数の回帰変数とし、以下の回帰モデルを考えよう。

$$
y_i = \beta_0 + \beta_1 D_i + u_i. \tag{43}
$$

ここで $E(u_i|D_i)=0$ とすると、

- $E(y_i|D_i=0)=\beta_0 \Rightarrow \beta_0$ は、$D_i=0$ の時の y_i の平均である。
- $E(y_i|D_i=1)=\beta_0+\beta_1 \Rightarrow \beta_0+\beta_1$ は、$D_i=1$ の時の y_i の平均である。
- β_1 はこれらの平均の差である。

β_1 の OLS 推定量は、これらの二つのグループから計算した標本平均の差に等しくなる。

多項回帰モデル　多項回帰モデルとは

$$
y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_r X_i^r + u_i \tag{44}
$$

のようなモデルである。r を次数といい、回帰に含まれている X の最大の乗数である。

- $r=2$ なら、2 次回帰モデル
- $r=3$ なら、3 次回帰モデルである。

回帰関数が線形かどうか調べるには、$H_0: \beta_2 = 0, \ldots, \beta_r = 0$ という帰無仮説を F 統計量を使って検定すればよい。

多項回帰モデルは、回帰変数から被説明変数への影響が、回帰変数の値に依存するという状況をモデル化する一つのやり方である。例として 2 次モデルを考えよう。

$$\begin{aligned} y &= \beta_0 + \beta_1 X + \beta_2 X^2 + u, & (45)\\ y + \Delta y &= \beta_0 + \beta_1 (X + \Delta X) + \beta_2 (X + \Delta X)^2 + u. & (46) \end{aligned}$$

このとき、

$$\Delta y = \beta_1 \Delta X + 2\beta_2 X \Delta X + \beta_2 (\Delta X)^2, \tag{47}$$

あるいは、

$$\frac{\Delta y}{\Delta X} = \beta_1 + 2\beta_2 X + \beta_2 \Delta X. \tag{48}$$

X の変化が y に与える影響は、

$$\widehat{\Delta y} = \hat\beta_1 \Delta X + 2\hat\beta_2 X \Delta X + \hat\beta_2 (\Delta X)^2 \tag{49}$$

として推定できる。注意すべきは、この影響は、X の初期値と変化の大きさ ΔX に依存していることである。

推定された影響の標準誤差も計算することができる。例えば、X を 10 から 11 に増やした時の y の変化分は、

$$\begin{aligned} \widehat{\Delta y} &= (\hat\beta_0 + \hat\beta_1 \times 11 + \hat\beta_2 \times 11^2) - (\hat\beta_0 + \hat\beta_1 \times 10 + \hat\beta_2 \times 10^2) & (50)\\ &= \hat\beta_1 + 21\hat\beta_2 & (51) \end{aligned}$$

である。したがって、$\widehat{\Delta y}$ の標準誤差は $SE(\hat\beta_1 + 21\hat\beta_2)$ であり、それは

$$SE(\Delta y) = \sqrt{\frac{1}{n}\begin{pmatrix} 0 & 1 & 21 \end{pmatrix} \hat V \begin{pmatrix} 0 \\ 1 \\ 21 \end{pmatrix}} \tag{52}$$

として計算できる。

- 言うまでもなく、β_1 を、X^2 を固定した時の、X が y に与える影響と考えるのはナンセンスである。

対数を使った回帰　y あるいは X の対数を使ったモデルもよく用いられる。対数の単位当たりの変化は、割合の変化として解釈できる。経済分析においては、変数の値そのものの変化よりも、割合の変化のほうに興味があることも多い。

例:

- 賃金格差。
- 所得の変化としては、10 万円の変化よりも 1%の変化に興味があることもある。
- ある変数の値を 1%変えた時の、他の変数の%で表した変化の大きさを、“弾力性” という。

$\Delta x/x$ が十分に小さいとき、対数と割合には次の関係が成り立つ。

$$\ln(x+\Delta x) - \ln(x) \approx \frac{\Delta x}{x}, \tag{53}$$

つまり、$\ln(x)$ が 0.01 変わったなら、それは、x が 1%変わったのと大体同じになる。

- 線形対数モデル: $y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$.
  - X の 1%の変化は、y を $0.01\beta_1$ 分増加させる。
- 対数線形モデル: $\ln(y_i) = \beta_0 + \beta_1 X_i + u_i$.
  - X を 1 単位変えた時、y は $100\beta_1$%だけ増加する。
- 対数対数モデル: $\ln(y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$.
  - X を 1%変えると、y は β_1%だけ変わる。係数 β_1 は y の X に対する "弾力性" である。

回帰変数の相互作用　相互作用をあらわす項をモデルに入れることによって、ある回帰変数の限界効果が、他の変数の値に依存する状況を表現することができる。

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i. \tag{54}$$

交差項 $X_{1i} \times X_{2i}$ は相互作用項と呼ばれる。X_2 を固定したときに、X_1 の変化が y に与える影響は X_2 の関数になっている。

$$\frac{\Delta y}{\Delta X_1} = \beta_1 + \beta_3 X_2. \tag{55}$$

係数 β_3 は、X_1 と X_2 を両方を同時に単位分だけ変化させたときの影響から、それぞれ片方だけ変化させたときの影響の和を引いた量であることがわかる。X_1 を ΔX_1 だけ変化させ、X_2 を ΔX_2 だけ変化させると、

$$\Delta y = (\beta_1 + \beta_3 X_2)\Delta X_1 + (\beta_2 + \beta_3 X_1)\Delta X_2 + \beta_3 \Delta X_1 \Delta X_2. \tag{56}$$

回帰変数が 2 項変数の場合にも同様のモデルが考えられるが、その時には解釈が少し異なる。

$$y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i. \tag{57}$$

- X_i: 連続な回帰変数。
- D_i: 2 項変数。

X の変化の y への影響は、

$$\begin{cases} \beta_1 & \text{もし } D = 0, \\ \beta_1 + \beta_3 & \text{もし } D = 1. \end{cases} \tag{58}$$

- この回帰は、D の値によって標本を二つに分け、それぞれの標本ごとに y を X に回帰するのと同じである。

また、回帰変数が 2 つとも 2 項変数である場合を考えることもできる。

$$y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 (D_{1i} \times D_{2i}) + u_i. \tag{59}$$

このモデルの場合、

$$\begin{aligned}
E(y_i|D_{1i} = 0, D_{2i} = 0) &= \beta_0 & (60)\\
E(y_i|D_{1i} = 1, D_{2i} = 0) &= \beta_0 + \beta_1 & (61)\\
E(y_i|D_{1i} = 0, D_{2i} = 1) &= \beta_0 + \beta_2 & (62)\\
E(y_i|D_{1i} = 1, D_{2i} = 1) &= \beta_0 + \beta_1 + \beta_2 + \beta_3 & (63)
\end{aligned}$$

となる。またこの回帰モデルの OLS 推定値は、標本を D_1 と D_2 の値によって 4 つにわけ、それぞれの標本の標本平均を計算したものことと同じになる。なお β_3 は、政策評価分析における差分内差分法で興味対象とされるパラメータである。

2.9 因果関係

2.9.1 因果、相関と仮定 1

$\hat{\beta}$ の推定量の形（X と Y の共分散/X の分散）からもわかるように、回帰モデルの推定からわかるのは、基本的に相関関係である。しかし、計量経済分析においては、暗に右辺の X が左辺の Y の決定要因である、言い換えると X から Y への因果関係があると考えることが多い。

どのような場合にそのように考えてよさそうか、単回帰モデル

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{64}$$

について仮定 1（$E(u|X) = 0$）と関連付けて例示しよう。例えば、ある肥料が農作物の収穫量にどのくらい影響を与えるか調べたいとしよう。面積当たりの肥料の量を X_i、収穫量を Y_i とする。u_i は日照時間や気温、元々の土地の肥沃度など、収穫量を決める要因である。肥料の効果を正確に調べたい研究者は、各栽培地ごとの肥料の量を u_i と全く無関係に変えて収穫量のデータを集めるだろう。つまり、X_i と u_i は独立になり、その結果仮定 1 が成り立つから、X_i が変わっても平均的には u_i は変化しない。従って X_i が 1 増えれば Y_i は平均的に β_1 だけ増える。このような状況なら、X_i の変化に起因する Y_i の平均的な変化を定量的に読み取ることができる。

さて、肥料の研究者の実験データではなく、実際の農家が与えた肥料の量と収穫量のデータを得て、そこから両者の関係を調べることを考えてみよう。農家はどのように肥料を与えるだろうか？u_i を日当たりと観測誤差を合わせたものとしよう。日当たりがよい土地をもつ農家はあまり肥料を使わなくてもすむが、日当たりの悪い土地で栽培する農家は多めに肥料を入れるだろう。日光の影響が大きい作物については、極端な場合には（日当たりが悪いために）肥料をたくさん使う土地からの収穫量の方が少ないという逆転現象さえ起こるかもしれない。このような事情を考慮せずにデータを眺めると、「この肥料は収穫量に悪影響を及ぼす」という誤った結果を導くことになってしまう。逆転まではしなくても、肥料の純粋な効果を正しく測ることができないことは明らかだろう。このデータについては、X_i と u_i に関係があるために仮定 1 は成り立たない。この状況をもう少し明示的に式で表すと、次のようになる。日当たりを w_i、観測誤差を v_i、$u_i = w_i + v_i$ としよう。農家は w_i に応じて X_i を決めており、その関係を $X_i = h(w_i)$ としよう。この関係を逆に表すと $w_i = k(X_i)$ と書ける。つまり、k は h の逆関数である。$u_i = k(X_i) + v_i$ なので

$$Y_i = \beta_0 + \beta_1 X_i + k(X_i) + v_i \tag{65}$$

となる。v_i は観測誤差だから、X_i が変化しても平均的には v_i は変化しない。従って、X_i の変化がもたらす平均的な Y_i の変化は $\beta_1 X_i$ と $g(X_i)$ の変化の合計になることがわかる。そのため、(60) 式から X_i の変化を原因とする Y_i の変化を正しく測れないのである。

以上のことから、仮定 1 が成り立たない場合には回帰モデルが因果関係を表さないことがわかるだろう。その意味で、計量経済分析では仮定 1 が成り立つ時には、(60) 式が因果関係をとらえていると解釈する立場もある。

また、より根本的な解決を目指して、そもそも仮定 1 が成り立つようにデータを集める工夫ができる場合もある。自然科学でよく行われるランダム化比較実験 (RCT; randomized controlled tiral) がその代表的なものである。しかし、自然科学と異なり、経済学においては社会的コストがかかり過ぎたり、倫理的な問題が生じうるため RCT は簡単ではない。例えば、もし日銀が為替介入の効果を調べるためにランダムなタイミングでランダムな金額の為替介入を行う実験をすれば、実体経済に大混乱をもたらすだろう。しかし、実験に近い状況になっていると考えられるデータを探して分析したり、実際に小規模な実験を行うなど、そのようなアプローチの研究も近年精力的に進められている。

2.9.2 因果関係を調べるモデル (Rubin の潜在効果モデル)

因果効果を統計学的にとらえるためのモデルは Rubin によって提案された。ここでは、最も基礎的なものを紹介する。

ある政策、介入や治療などを取った時に、それが原因でターゲットとなる変数にどれだけの影響があったかを調べるのが因果効果分析である。経済学の文脈では、例えば職業訓練を受けた時に、その結果として就職確率や賃金がどれだけ変化するかが因果効果である。実際のデータでは例えば訓練前の賃金と訓練後の賃金が観察されるが、仮に上昇していたとしても、実はマクロ経済状態の改善やその他の状況も含まれている可能性があり、訓練そのものの直接効果を調べることは単純ではない。因果効果モデルでは、それらの外部的な要因も明示的に扱ってできるだけ直接的効果を正しく計測することを目指す。

ルービンの潜在効果モデル（potential outcome model） 以下では、職業訓練の例で説明する。個人 i が職業訓練を受けるかどうかのダミー変数を D_i とし、

$$D_i = \begin{cases} 1, & \text{職業訓練を受けた} \\ 0, & \text{受けていない} \end{cases} \tag{66}$$

とする。さらに、訓練を受けたときと受けなかったときの潜在的結果（例えば賃金）を

$$Y_i = \begin{cases} Y_i(1), & D_i = 1 \\ Y_i(0), & D_i = 0 \end{cases} \tag{67}$$

何を推定したい？ 個人 i にとっての因果効果は $\tau_i = Y_i(1) - Y_i(0)$ で定義される。しかし、個人 i は訓練を受けるか受けないかのどちらかで、データとして観察されるのは $Y_i(1)$ と $Y_i(0)$ のどちらか一方のみであるため、実際にはそれは計算できない。推定する量として最も初歩的なものは平均処置効果（ATE:Average Treatment Effect）と処置群に対する平均処置効果（ATT: Average Treatment Effect for Treated）である。それらはそれぞれ、

$$ATE : \tau_{ATE} = E(Y_i(1) - Y_i(0)) \tag{68}$$

$$ATT : \tau_{ATT} = E(Y_i(1) - Y_i(0)|D_i = 1) \tag{69}$$

である。更に、もし共変量 X_i や二項変数 Z_i が観察されて $D_i = D_i(Z_i)$ の場合、

$$\begin{aligned} CATE(x) &= E(Y_i(1) - Y_i(0)|X_i = x) && (70) \\ LATE &= E(Y_i(1) - Y_i(0)|D_i(1) > D_i(0)) && (71) \\ & && (72) \end{aligned}$$

などが推定対象である。他にも状況や関心によって様々な量が考えられる。

ここでは ATE の推定について解説する。

2.9.3 ATE の推定

無作為標本 $(Y_1, D_1), \cdots, (Y_n, D_n)$ があるとする。ここで観測される賃金は

$$Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0), i = 1, \cdots .n \tag{73}$$

と表すことができる。この表現を用いると、

$$\begin{aligned} E(Y_i|D_i = 1) &= E(D_i Y_i(1) + (1 - D_i)Y_i(0)|D_i = 1) && (74) \\ &= E(Y_i(1)|D_i = 1) && (75) \end{aligned}$$

$$\begin{aligned} E(Y_i|D_i = 0) &= E(D_i Y_i(1) + (1 - D_i)Y_i(0)|D_i = 0) && (76) \\ &= E(Y_i(0)|D_i = 0) && (77) \end{aligned}$$

が成り立つことが分かる。

1) $(Y_i(0), Y_i(1)) \perp D_i$ **のとき（RCT）**　これは訓練の割り当てが潜在的結果とは無関係でランダムと仮定である。(74)-(77) とこの仮定の下で、

$$E(Y_i|D_i = 1) = E(Y_i(1)|D_i = 1) = E(Y_i(1)) \tag{78}$$

$$E(Y_i|D_i = 0) = E(Y_i(0)|D_i = 0) = E(Y_i(0)) \tag{79}$$

が成り立つので

$$\begin{aligned} \tau_{ATE} &= E(Y_i(1) - Y_i(0)) && (80) \\ &= E(Y_i|D_i = 1) - E(Y_i|D_i = 0) && (81) \end{aligned}$$

となるが、これは訓練を受けた人の賃金の平均 $(\bar{Y}(1))$ から訓練を受けなかった人の賃金の平均 $(\bar{Y}(0))$ を引いたもので推定できる。つまり、

$$\hat{\tau}^{RCT}_{ATE} = \bar{Y}(1) - \bar{Y}(0) \tag{82}$$

同じ結果を回帰モデルの推定をベースに導出可能である。(73) を変形すると、

$$\begin{aligned} Y_i &= Y_i(0) + \{Y_i(1) - Y_i(0)\}D_i && (83) \\ &= E\{Y_i(0)\} + E\{Y_i(1) - Y_i(0)\}D_i + u_i && (84) \end{aligned}$$

となる。ただし、$u_i = [Y_i(0) - EY_i(0)] + [(Y_i(1) - Y_i(0)) - E\{Y_i(1) - Y_i(0)\}]D_i$ である。したがって、(84) で $E(u_i|D_i) = 0$ なら OLS によって $E\{Y_i(1) - Y_i(0)\}$ の一致推定量が得られる。上の仮定はそのための十分条件になっている。実際、

$$\begin{aligned} \tilde{\tau}^{RCT}_{ATE} &= \{\sum_{i=1}^{n}(D_i - \bar{D})^2\}^{-1}\sum_{i=1}^{n}(D_i - \bar{D})Y_i && (85) \\ &= \bar{Y}(1) - \bar{Y}(0) = \hat{\tau}^{RCT}_{ATE} && (86) \end{aligned}$$

2) RCT でない場合　RCT でない場合にそれを無視して (82) を計算しても、τ_{ATE} を正しく推定できない。

研究者がデータ収集をデザインできる実験研究では RCT の仮定が満たされるデータが得られるが、社会科学で見られる観察データではその条件は満たされない。典型的なのは、訓練参加（D_i）と潜在的結果変数 $(Y_i(1), Y_i(0))$ の両方に影響を与える共通の共変量 X_i（交絡変数という）があるといった状況である。例えば訓練前の収入が低い人は訓練に参加しやすいが、（もしかすると仕事の能力が高くないために）潜在的結果の賃金は低いかもしれない。そのような状況では、交絡変数 X_i をコントロールすることによって訓練参加と潜在的結果変数が独立になる場合、ATE をうまく推定でき、以下では 3 つの方法を解説する。

まず、二つの条件を仮定する。第一は「条件付き独立性（非交絡）」で、

$$(Y_i(0), Y_i(1)) \perp D_i | X_i \tag{87}$$

第二は「オーバーラップ」の仮定で、任意の x に対してが成り立つとする。

$$0 < P(D_i = 1 | X_i = x) < 1 \tag{88}$$

この二つの仮定を合わせて「強い無視可能性 (Strong ignorability)」ともいう。

(1) 回帰調整法（Regression Adjustment）　ATE の推定のために、CATE(conditional ATE) と呼ばれる以下の量を考える。

$$\tau_{CATE}(X_i) = E\{Y_i(1) - Y_i(0) | X_i\} \tag{89}$$

これを用いると、繰り返し期待値の法則より $\tau_{ATE} = E\{\tau_{CATE}(X_i)\}$ であるから、$\tau_{CATE}(\cdot)$ がわかれば $\sum_{i=1}^{n} \tau_{CATE}(X_i)/n$ により ATE が推定される。

条件付き独立の仮定から、

$$\begin{aligned} m_1(x) &\equiv E\{Y_i(1)|X_i = x\} = E\{Y_i(1)|X_i = x, D_i = 1\} = E\{Y_i|X_i = x, D_i = 1\} && (90) \\ m_0(x) &\equiv E\{Y_i(0)|X_i = x\} = E\{Y_i(0)|X_i = x, D_i = 0\} = E\{Y_i|X_i = x, D_i = 0\} && (91) \end{aligned}$$

なので、$D_i = 1$ と $D_i = 0$ それぞれのグループごとに Y_i を X_i に回帰すればこれらが推定でき、その差を計算すればよい。最も単純なやり方は、

$$\begin{aligned} Y_i &= \alpha_1 + \alpha_1 X_i + \epsilon_{1i} : (D_i = 1) && (92) \\ Y_i &= \alpha_0 + \alpha_0 X_i + \epsilon_{0i} : (D_i = 0) && (93) \end{aligned}$$

という二つの回帰推定を行って

$$\begin{aligned} \hat{m}_1(X_i) &= \hat{\alpha}_1 + \hat{\alpha}_1 X_i && (94) \\ \hat{m}_0(X_i) &= \hat{\alpha}_0 + \hat{\alpha}_0 X_i && (95) \end{aligned}$$

を計算し、そこから

$$\hat{\tau}_{ATE} = \frac{1}{n} \sum_{i=1}^{n} \{\hat{m}_1(X_i) - \hat{m}_0(X_i)\} \tag{96}$$

によって ATE を計算する。

(2) 傾向スコアマッチング (Propensity score matching)　$p(x) = P(D_i = 1|X_i = x)$ を傾向スコアという。$D_iY_i = D_iY_i(1)$ は明らかなので、

$$E\{\frac{D_iY_i}{p(X_i)}\} = E\{\frac{E(D_iY_i|X_i)}{p(X_i)}\} = E\{\frac{E(D_iY_i(1)|X_i)}{p(X_i)}\} \quad (97)$$

$$= E\{\frac{E(D_i|X_i)E(Y_i(1)|X_i)}{p(X_i)}\} = E\{\frac{p(X_i)E(Y_i(1)|X_i)}{p(X_i)}\} \quad (98)$$

$$= E[E\{Y_i(1)|X_i\}] = E\{Y_i(1)\} \quad (99)$$

が成り立つ。ここでオーバーラップの仮定が用いられている。全く同様にして、

$$E\{\frac{(1-D_i)Y_i}{1-p(X_i)}\} = E\{Y_i(0)\} \quad (100)$$

を示すことができる。以上の計算から、ATE は

$$\tau_{ATE} = E\{\frac{D_iY_i}{p(X_i)}\} - E\{\frac{(1-D_i)Y_i}{1-p(X_i)}\} \quad (101)$$

と表現することができ、それを用いて以下のように推定ができる。

1．傾向スコア $p(x)$ を推定し（後に学ぶプロビットやロジットモデルが使われる）、それを $\hat{p}(x)$ とする。

2．それを用いて

$$\hat{\tau}_{ATE}^{PSM} = \frac{1}{n}\sum_{i=1}^{n}\{\frac{D_iY_i}{\hat{p}(X_i)} - \frac{(1-D_i)Y_i}{1-\hat{p}(X_i)}\} \quad (102)$$

によって ATE を推定する。

この推定では $p(x)$ の逆数が用いられるため、inverse probability weighting 法とも呼ばれる。

(3) 二重頑健推定法 (Doubly robust estimation)　回帰調整法と傾向スコアマッチング法ではそれぞれ $m_1(x), m_0(x)$ と $p(x)$ の推定が求められ、多くの場合は線形モデルなど適当なモデルを当てはめる。モデルが正しくて、それらを一致推定できればこの方法は正しい ATE の推定値を与えるが、間違っていると一致性を失う。Doubly robust 法は、$(m_1(x), m_0(x))$ か $p(x)$ のどちらかが正しく推定されていれば正しく ATE を推定できる方法である。まず、$p(x) = P(D_i = 1|X_i = x)$ なので、繰り返し期待値の法則から

$$E\{(1-\frac{D_i}{p(X_i)})m_1(X_i)\} = E\{(1-\frac{E(D_i|X_i)}{p(X_i)})m_1(X_i)\} = 0 \quad (103)$$

$$E\{(1-\frac{1-D_i}{1-p(X_i)})m_0(X_i)\} = E\{(1-\frac{1-E(D_i|X_i)}{1-p(X_i)})m_0(X_i)\} = 0 \quad (104)$$

が成り立つ。したがって、$(\hat{m}_1(x), \hat{m}_0(x))$ と $\hat{p}(x)$ が一致推定量なら、(102) の $\hat{\tau}_{ATE}^{PSM}$ の右辺にそれらのサンプル対応を加えて

$$\hat{\tau}_{ATE}^{DR} = \frac{1}{n}\sum_{i=1}^{n}[\frac{D_iY_i}{\hat{p}(X_i)} - (1-\frac{D_i}{\hat{p}(X_i)})\hat{m}_1(X_i) \quad (105)$$

$$- \{\frac{(1-D_i)Y_i}{1-\hat{p}(X_i)} - (1-\frac{1-D_i}{1-\hat{p}(X_i)})\hat{m}_0(X_i)\}] \quad (106)$$

としても ATE は一致推定されることは明らかである。

実は、もし $\hat{m}_1(x), \hat{m}_0(x)$ または $\hat{p}(x)$ のどちらかが一致推定量でなくても $\hat{\tau}_{ATE}^{DR}$ で ATE が正しく推定されることを確認しよう。まず、$\hat{m}_1(x), \hat{m}_0(x)$ が間違った関数 $g_1(x), g_0(x)$ に収束する場合、(105)-(106) の第２、４項はそれぞれ

$$
\begin{aligned}
E\{(1-\frac{D_i}{p(X_i)})g_1(X_i)\} &= 0 && (107)\\
E\{(1-\frac{1-D_i}{1-p(X_i)})g_0(X_i)\} &= 0 && (108)
\end{aligned}
$$

に収束することがわかり、そのためバイアスは生じない。他方、$\hat{p}(x)$ が間違った関数 $q(x)$ に収束するとき、$\hat{\tau}_{ATE}^{DR}$ の１行目と２行目は、繰り返し期待値の法則より、それぞれ

$$
\begin{aligned}
E[\frac{D_iY_i}{q(X_i)}-(1-\frac{D_i}{q(X_i)})m_1(X_i)] &= E\{m_1(X_i)\}+E\{\frac{D_i(Y_i-m_1(X_i))}{q(X_i)}\} && (109)\\
&= E\{m_1(X_i)\} && (110)\\
E[\frac{(1-D_i)Y_i}{1-q(X_i)}-(1-\frac{1-D_i}{1-q(X_i)})m_0(X_i)\}] &= E\{m_0(X_i)\}+E\{\frac{(1-D_i)(Y_i-m_0(X_i))}{1-q(X_i)}\} && (111)\\
&= E\{m_0(X_i)\} && (112)
\end{aligned}
$$

に収束する。したがって、$p(x)$ の推定が間違っていても $\hat{m}_1(x), \hat{m}_0(x)$ が一致推定量なら、やはり ATE は正しく推定される。