2025/07/11

藪友良

クラスターロバスト標準誤差

パネルデータでは、個体iに着目すれば誤差項$e_{i,t}$は計T個の時系列データ$(e_{i,1}, e_{i,2}, \dots, e_{i,T})$となります。この$T$個の時系列データは、1 つのクラスター(集団)を形成し、その中で系列相関が発生しています。推定量の標準誤差は、こうしたクラスター構造に対して頑健な標準誤差を用いる必要があります。この標準誤差は、**クラスターロバスト標準誤差**(cluster robust standard error)と呼ばれます。クラスターロバスト標準誤差は、**不均一分散と系列相関に頑健な標準誤差(HAC 標準誤差)**の 1 種となります。ここでは、クラスターロバスト標準誤差の考え方と推定方法を紹介します。

本稿は、藪友良『入門 実践する計量経済学』(2023 年、東洋経済新報社)の補足資料です。文中に参照節が書かれていますが、それはこちらの本の節となります。

1. プールド OLS

個別要因Z_iが存在しない場合、モデルは次のようになります。

$$Y_{i,t} = \alpha + \beta X_{i,t} + u_{i,t}$$

個別要因がないため定数項は一定としています。個体は i=1,2,.,,N、時点は t=1,2,…T ですから、サンプルサイズは $N\,T$ です。個体iとjが異なるなら、どの時点tとsに対しても誤差項は無相関とします。

$$E\left[u_{i,t}u_{j,s}\right] = 0$$

しかし、誤差項は個体内(クラスター内)では、系列相関が許容されます。

$$E\left[u_{i,t}u_{i,s}\right] \neq 0$$

ここでは、プールド OLS 推定量におけるクラスター標準誤差の推定方法を説明します。

1.1. プールド OLS 推定量 の分散

プールド OLS 推定量は、単回帰による通常の OLS と同じで次のように表せます(OLS 推定量の公式は P32 参照)[1]。

$$\hat{\beta} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{i,t} - \bar{X})\,(Y_{i,t} - \bar{Y})}{\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{i,t} - \bar{X})^2}$$

---

[1] ここで分母の 2 重和は次のように展開できます(分子も同様に展開できます)。

$$\begin{aligned}
\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{i,t} - \bar{X})^2 &= \sum_{i=1}^{N}\left(\sum_{t=1}^{T}(X_{i,t} - \bar{X})^2\right) \\
&= \sum_{i=1}^{N}\left((X_{i,1} - \bar{X})^2 + (X_{i,2} - \bar{X})^2 + \cdots + (X_{i,T} - \bar{X})^2\right) \\
&= \left((X_{1,1} - \bar{X})^2 + (X_{1,2} - \bar{X})^2 + \cdots + (X_{1,T} - \bar{X})^2\right) \\
&\quad + \left((X_{2,1} - \bar{X})^2 + (X_{2,2} - \bar{X})^2 + \cdots + (X_{2,T} - \bar{X})^2\right) \\
&\quad \cdots \\
&\quad + \left((X_{N,1} - \bar{X})^2 + (X_{N,2} - \bar{X})^2 + \cdots + (X_{N,T} - \bar{X})^2\right)
\end{aligned}$$

ここで、その確率的表現は次のようになります(確率的表現は教科書 P58 参照)。

$$\hat{\beta} = \beta + \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{i,t} - \bar{X})\, u_{i,t}}{\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{i,t} - \bar{X})^2}$$

$$= \beta + \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} v_{i,t}}{\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{i,t} - \bar{X})^2}$$

式展開では、教科書 10.3.1 節と同様に、$v_{i,t} = (X_{i,t} - \bar{X})u_{i,t}$と定義しました。

ここで推定量$\hat{\beta}$の分散は、$X_{i,t}$が非確率変数とすると、次のようになります。

$$E\left[\left(\hat{\beta} - \beta\right)^2\right] = E\left[\left(\frac{\sum_{i=1}^{N}\sum_{t=1}^{T} v_{i,t}}{\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{i,t} - \bar{X})^2}\right)^2\right]$$

$$= \frac{E\left[\left(\sum_{i=1}^{N} q_i\right)^2\right]}{\left\{\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{i,t} - \bar{X})^2\right\}^2}$$

最後の式展開では、$q_i = \sum_{t=1}^{T} v_{i,t}$と定義しました。上式分子は、$\sigma_q^2 = E[q_i^2]$を仮定すると[2]、

$$E\left[\left(\sum_{i=1}^{N} q_i\right)^2\right] = E[q_1^2] + E[q_2^2] + \cdots + E[q_N^2] = N\sigma_q^2$$

となります(ここで、$E\left[u_{i,t}u_{j,s}\right] = 0$から$E\left[q_i q_j\right] = 0$)[3]。

以上の結果から、推定量$\hat{\beta}$の分散は次のように表せます。

$$E\left[\left(\hat{\beta} - \beta\right)^2\right] = \frac{N\sigma_q^2}{\left\{\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{i,t} - \bar{X})^2\right\}^2}$$

ここで、$\hat{\beta}$の分散の推定には、σ_q^2の推定が必要となることがわかります。

$$\sigma_q^2 = E[q_i^2] = E\left[\left(\sum_{t=1}^{T} v_{i,t}\right)^2\right]$$

1.2. σ_q^2の推定

ここで、σ_q^2は自己共分散$E[v_{i,t}v_{i,s}]$から構成されます。10.3 節では、T が大きいとして、これらの自己共分散を標本自己共分散によって推定しました。しかし、パネルデータでは、一般に Tが小さく、この方法を用いることはできません。

パネルデータは一般に、T は小さいですが、N が大きいという特徴を使います。具体的

---

[2] 読者の中には、この仮定に違和感を持たれた方もいるかと思います。しかし、この仮定は説明変数X_tが非確率変数ではなく、確率変数とすれば現実的仮定です。

[3] $E\left[u_{i,t}u_{j,s}\right] = 0$を用いると、

$$E[q_i q_j] = E\left[\left(\sum_{t=1}^{T} v_{i,t}\right)\left(\sum_{t=1}^{T} v_{j,t}\right)\right] = E\left[\left(\sum_{t=1}^{T}(X_{i,t} - \bar{X})u_{i,t}\right)\left(\sum_{t=1}^{T}(X_{j,t} - \bar{X})u_{j,t}\right)\right] = 0$$

には、プールド OLS からの残差$\hat{u}_{i,t}$を用いて、$\hat{q}_i = \sum_{t=1}^{T} \hat{v}_{i,t} = \sum_{t=1}^{T}(X_{i,t} - \bar{X})\hat{u}_{i,t}$を計算します。そして、これを各個体について計算し、

$$\hat{q}_1, \hat{q}_2, \dots, \hat{q}_N$$

これらの 2 乗の平均として、$\sigma_q^2 = E[q_i^2]$を推定します。

$$\hat{\sigma}_q^2 = \frac{1}{N}\sum_{i=1}^{N} \hat{q}_i^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=1}^{T} \hat{v}_{i,t}\right)^2$$

Nが十分に大きいなら、$\hat{\sigma}_q^2$はσ_q^2を正確に推定できるはずです。

分散の式に推定量$\hat{\sigma}_q^2$を代入すると、分散の推定量が得られます。

$$\frac{N\hat{\sigma}_q^2}{\left\{\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{i,t} - \bar{X})^2\right\}^2} = \frac{\sum_{i=1}^{N}\left(\sum_{t=1}^{T} \hat{v}_{i,t}\right)^2}{\left\{\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{i,t} - \bar{X})^2\right\}^2}$$

分散の推定量の平方根が**クラスターロバスト標準誤差**になります。

$$\sqrt{\frac{\sum_{i=1}^{N}\left(\sum_{t=1}^{T} \hat{v}_{i,t}\right)^2}{\left\{\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{i,t} - \bar{X})^2\right\}^2}}$$

2. 固定効果推定量

個別要因Z_iがある場合、モデルは次のようになります。

$$Y_{i,t} = \alpha_i + \beta X_{i,t} + u_{i,t}$$

ここで定数項は個体によって異なります。誤差項に関しては、これまでと同様、クラスター間では無相関ですが、クラスター内では系列相関があると仮定します。ここでは、個別要因が存在するため、固定効果推定を行います。

2.1.固定効果推定量の別表現

各個体について時間平均との差をとることで個別要因を取り除く作業を行います。まず、時間に関して和をとります。

$$\sum_{s=1}^{T} Y_{i,s} = \sum_{s=1}^{T}\left(\alpha_i + \beta X_{i,s} + u_{i,s}\right)$$

$$= T\alpha_i + \beta\sum_{s=1}^{T} X_{i,s} + \sum_{s=1}^{T} u_{i,s}$$

上式の両辺をTで割ると、時間平均の関係式が得られます。

$$\bar{Y}_i = \alpha_i + \beta\bar{X}_i + \bar{u}_i$$

ただし、時間平均は次のように定義しました。

$$\bar{Y}_i = \frac{1}{T}\sum_{s=1}^{T} Y_{i,s}、\bar{X}_i = \frac{1}{T}\sum_{s=1}^{T} X_{i,s}、\bar{u}_i = \frac{1}{T}\sum_{s=1}^{T} u_{i,s}$$

次に、$Y_{i,t}$から時間平均$\bar{Y}_i$を引くと、次のようになります。

$$Y_{i,t} - \bar{Y}_i = \left(\alpha_i + \beta X_{i,t} + u_{i,t}\right) - (\alpha_i + \beta \bar{X}_i + \bar{u}_i)$$
$$= \beta\left(X_{i,t} - \bar{X}_i\right) + (u_{i,t} - \bar{u}_i)$$

ここで、時間平均との差をそれぞれ

$$\tilde{Y}_{i,t} = Y_{i,t} - \bar{Y}_i$$
$$\tilde{X}_{i,t} = X_{i,t} - \bar{X}_i$$
$$\tilde{u}_{i,t} = u_{i,t} - \bar{u}_i$$

と定義すると、個別要因が除去された次式が得られます。

$$\tilde{Y}_{i,t} = \beta \tilde{X}_{i,t} + \tilde{u}_{i,t}$$

ここで、定数項は 0 になっていますから、係数βは定数項なしの OLS 推定量として求めることができます。

$$\hat{\beta} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{i,t}\tilde{Y}_{i,t}}{\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{i,t}^2}$$

定数項なしの OLS 推定は、3 章練習問題 12 を参照してください。なお、上式は、11.3 節のダミー変数を用いた固定効果推定量と同じになることが知られています[4]。

2.2.固定効果推定量の分散

固定効果推定量の確率的表現を求めます。ここで、$\tilde{Y}_{i,t} = \beta \tilde{X}_{i,t} + \tilde{u}_{i,t}$を代入すると、次のようになります。

$$\hat{\beta} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{i,t}\tilde{Y}_{i,t}}{\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{i,t}^2} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{i,t}\left(\beta\tilde{X}_{i,t} + \tilde{u}_{i,t}\right)}{\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{i,t}^2}$$
$$= \beta + \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{i,t}\tilde{u}_{i,t}}{\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{i,t}^2}$$
$$= \beta + \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}v_{i,t}}{\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{i,t}^2}$$

最後の式展開では、$v_{i,t} = \tilde{X}_{i,t}\tilde{u}_{i,t}$と定義しました。

固定効果推定量$\hat{\beta}$の分散は、$X_{i,t}$が非確率変数とすると、次のようになります。

$$E\left[\left(\hat{\beta} - \beta\right)^2\right] = E\left[\left(\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}v_{i,t}}{\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{i,t}^2}\right)^2\right]$$
$$= \frac{E\left[\left(\sum_{i=1}^{N}q_i\right)^2\right]}{\left\{\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{i,t}^2\right\}^2}$$

式展開では、$q_i = \sum_{t=1}^{T}v_{i,t}$と定義しました。上式分子は$\sigma_q^2 = E[q_i^2]$を仮定すると、

---

[4] FWL 定理によって両者が同じになることを示すことができます。興味がある方は、本書サポートウェブサイトの「FWL 定理と重回帰分析」と 11 章模範解答の補足を参照してください。

$$E\left[\left(\sum_{i=1}^{N} q_i\right)^2\right] = E[q_1^2] + E[q_2^2] + \cdots + E[q_N^2] = N\sigma_q^2$$

となります。以上の結果から、推定量$\hat{\beta}$の分散は次のように表せます。

$$E\left[\left(\hat{\beta} - \beta\right)^2\right] = \frac{N\sigma_q^2}{\left\{\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{i,t}^2\right\}^2}$$

これはプールド OLS の分散と類似の結果であり、σ_q^2の推定も同じ方法をとることができます。

具体的には、固定効果推定の残差$\hat{u}_{i,t}$を用いて、$\hat{q}_i = \sum_{t=1}^{T}\hat{v}_{i,t} = \sum_{t=1}^{T}\tilde{X}_{i,t}\hat{u}_{i,t}$を計算します。そして、これら 2 乗の平均として、$\sigma_q^2 = E[q_i^2]$を推定します。

$$\hat{\sigma}_q^2 = \frac{1}{N}\sum_{i=1}^{N}\hat{q}_i^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\hat{v}_{i,t}\right)^2$$

そして、分散の式に推定量$\hat{\sigma}_q^2$を代入し平方根をとると**クラスターロバスト標準誤差**が得られます。

$$\sqrt{\frac{N\hat{\sigma}_q^2}{\left\{\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{i,t}^2\right\}^2}} = \sqrt{\frac{\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\hat{v}_{i,t}\right)^2}{\left\{\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{i,t}^2\right\}^2}}$$

3. まとめ

ここではパネルデータにおけるクラスターロバスト標準誤差を紹介しました。クラスターロバスト標準誤差は HAC 標準誤差の 1 種ですが、10 章で Tが∞になるとしたアプローチとは異なり、Tが小さく Nが∞になるとしたアプローチを用いています。Tが大きく Nが小さい場合、もしくは、誤差項がクラスター間で相関している場合については、Stock and Watson の『Introduction to Econometrics』を参照してください。

本稿では説明していませんが、クラスターロバスト標準誤差はパネルデータだけでなく、横断面データでも頻繁に用いられます。たとえば、小学生の標準テストの成績とクラスサイズの関係が知りたいとします。生徒は互いに影響を与え合うと考えると、生徒間の成績がクラス内で相関している可能性があります。この場合、1 クラスを 1 つのクラスター(集団)と考えて、クラス内での誤差項同士の相関を許容したクラスターロバスト標準誤差を用います。もし 1 つの学校内で相互の成績に相関があれば、学校を 1 つのクラスターとすることも可能です。どれをクラスターとするかは難しい問題であり、分析対象によって異なります。