2025/09/17

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 種です。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。なお、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。

1. プールド OLS

ここで次のモデルを考えましょう。

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

個別効果がないため、定数項は一定としました。個体は$i = 1,2,\dots,N$、時点は$t = 1,2,\dots,T$ですから、サンプルサイズは NT 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。

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

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。

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

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] 分母の 2 重和 ΣΣ は次のように展開できます。

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

つまり、これは$X_{i,t}$の偏差の 2 乗和になります。

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『入門 実践する計量経済学』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、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、

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

となります(式展開では、iとjが異なるとき、$E[q_i q_j] = 0$を用いました)[2]。

$\sigma_{qi}^2 = E[q_i^2]$と表記すると、推定量$\hat{\beta}$の分散は次のように表せます。

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

分散の平均$\bar{\sigma}_q^2 = \frac{1}{N}\sum_{i=1}^{N}\sigma_{qi}^2$と定義すると、推定量$\hat{\beta}$の分散は、次のようにも表現できます。

$$\frac{N\bar{\sigma}_q^2}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2}$$

---

[2] iとjが異なるとき、期待値$E[q_i q_j]$は 0 になります。

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

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

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

最後の展開では、iとjが異なるとき、$E[u_{i,t}u_{j,s}] = 0$となることを用いました。

1.2. クラスターロバスト標準誤差

ここで、σ_{qi}^2は定義から次のようになります。

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

つまり、σ_{qi}^2は自己共分散$E[v_{i,t}v_{i,s}]$から構成されるわけです。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。

ここで、$\sigma_{qi}^2 = E[q_i^2]$であるため、その推定量として$q_i^2 = (\sum_{t=1}^{T} v_{i,t})^2 = (\sum_{t=1}^{T}(X_{i,t} - \bar{X})u_{i,t})^2$を用いることが自然です。しかし、誤差項$u_i$が観察できないため、誤差項$u_i$の推定量である残差$\hat{u}_i$に置き換えることで、$\sigma_{qi}^2$の推定量$\hat{q}_i^2$を次のように求めることができます。

$$\hat{q}_i^2 = \left(\sum_{t=1}^{T} \hat{v}_{i,t}\right)^2 = \left(\sum_{t=1}^{T}(X_i - \bar{X})^2\hat{u}_i^2\right)^2$$

しかし、これはσ_{qi}^2を観測値$\hat{q}_i^2$だけで推定しており、その精度精度はかなり低くなります。

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があります。M.アレラーノ(Manuel Arellano)は、この特徴を生かして、個々のσ_{qi}^2ではなく、その平均$\bar{\sigma}_q^2$なら高い精度で推定できるとしました[3]。つまり、個々の$\hat{q}_i^2$は推定誤差が大きい一方、Nが十分に大きければ、その平均である

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

は個々の推定誤差が打ち消しあうことで正確に推定できるわけです。

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

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

そして、分散の推定量の平方根が**クラスターロバスト標準誤差**(cluster robust standard errors)になります。

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

---

[3] Arellano, Manuel. "Computing robust standard errors for within-groups estimators." *Oxford Bulletin of Economics & Statistics* 49.4 (1987).なお、Arellano のアイデアは、ロバスト標準誤差を開発した H.ホワイト(Halbert White)のアイデアに基づいています。詳しくは、サポートウェブサイトの「不均一分散に頑健な標準誤差」、もしくは以下の論文を読んでください。White, Halbert. "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity." *Econometrica: journal of the Econometric Society* (1980): 817-838.

実証分析では、クラスターロバスト標準誤差は、ロバスト標準誤差よりも大きな値をとることが多くなります。パネルデータでは、系列相関が生じることが一般的ですから、ぜひクラスターロバスト標準誤差を用いるようにしましょう。

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。

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

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。

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

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

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 章の練習問題 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. クラスターロバスト標準誤差

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。

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

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

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

ただし、式展開では、$q_i = \sum_{t=1}^{T}v_{i,t}$、$\sigma_{qi}^2 = E[q_i^2]$、$\bar{\sigma}_q^2 = \frac{1}{N}\sum_{i=1}^{N}\sigma_{qi}^2$としました。これはプールド OLS と類似の結果であり、$\bar{\sigma}_q^2$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、2 乗の平均として$\hat{\bar{\sigma}}_q^2$を推定します。

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

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

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

---

[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。

3. まとめ

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、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。

補足：横断面データとクラスターロバスト標準誤差

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ることが適切となります[5]。

ここで次のモデルを考えます。

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

jクラスの生徒iの点数を$Y_{i,j}$、そしてjクラスの生徒iの勉強時間を$X_{i,j}$とします。ここで、$j = 1,2,..,J$とし、また、$i = 1,2,\dots,N_j$とします。つまり、クラスは計J個あり、クラスjにはN_j人の生徒がいます。このため、サンプルサイズは$\sum_{j=1}^{J} N_j$となります。

このとき、OLS 推定量は

$$\hat{\beta} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{N_j}(X_{i,j} - \bar{X})\,(Y_{i,j} - \bar{Y})}{\sum_{j=1}^{J}\sum_{i=1}^{N_j}(X_{i,j} - \bar{X})^2}$$

このとき、その確率的表現は次のようになります。

$$\hat{\beta} = \beta + \frac{\sum_{j=1}^{J}\sum_{i=1}^{N_j}(X_{i,j} - \bar{X})\,u_{i,j}}{\sum_{j=1}^{J}\sum_{i=1}^{N_j}(X_{i,j} - \bar{X})^2}$$

これは 1.1 節と同じ式であり、同様の式展開をすることでクラスターロバスト標準誤差を計算することができます。

---

[5] たとえば、横断面データを Stata で分析するときは、reg Y X, cluster(class_id)とすれば、クラスターロバスト標準誤差を計算できます。ここで、Y は被説明変数、X は説明変数、class_id はクラス番号に当たります。クラスターロバスト標準誤差は、通常のロバスト標準誤差よりも大きな値をとることが多くなります。