福山平成大学経営学部紀要
第 16 号（2020），105－125 頁

社会システム分析のための統合化プログラム３６ －パネルデータ分析・操作変数回帰分析－

福井 正康[*1]・兎内 祥子[*1]

[*1] 福山平成大学経営学部経営学科

要旨：我々は教育分野での利用を目的に社会システム分析に用いられる様々な手法を統合化したプログラム College Analysis を作成してきた。現在、経済特有の考え方や分析手法を取り入れることを目標にプログラム開発を進めており、今回は回帰分析における欠落バイアスの除去を目的とした、パネルデータ分析と操作変数回帰分析を組み込んだ。いずれも初心者が簡単に試せることを目的に開発している。

キーワード：パネルデータ分析、操作変数回帰分析、College Analysis、統計

1. はじめに

経済の分野では、標本のデータ数が多く取れるという特徴から、検定やパラメータの推定で中心極限定理を仮定した分析が多く用いられる。また、因果関係の説明や予測に回帰分析が利用されることが多く、欠落変数のバイアス除去のために、様々な拡張された形式の回帰分析が考えられている。我々が開発してきた分析ソフト College Analysis は元々保健分野からの出発であったため、経済分野の分析には十分といえない部分があった。我々はこの状況を改善すべく、経済分野の学習者も気軽に分析を試すことができるように、計量経済で利用される代表的な手法を College Analysis の中で開発していくことにした。この論文では、個体効果や時系列効果があるデータからそれらの影響を取り除く手法であるパネルデータ分析と、除外された観測されないバイアス要因から影響を受ける説明変数から、その影響を取り除く手法である操作変数回帰分析を取り上げる。これらの分析手法は保健分野ではなじみの薄い手法である。最後に、この論文で利用される、回帰分析に関する行列と確率分布の公式について補遺にまとめておく。本文の中からは【公式1】などのように引用する。

2. パネルデータ分析

重回帰分析で、除外された変数がある場合、時系列データが与えられていると、その影響を取り除くことが可能になることがある。例えばそれが個体に依存し、時間には依存しない固定効果の場合、各個体の測定値の時間平均を引くと固定効果を消し去ることができる。またそれが時間に依存し、個体には依存しない時間効果の場合、各時間の測定値の個

体平均を引くと時間効果を消し去ることができる。パネルデータ分析は、この性質を利用して個体や時間に依存する特殊な影響を取り除いて重回帰分析を行う手法である。

パネルデータ分析は、例えば固定効果だけの場合、以下のような形の回帰分析である。

$$y_{it} = \sum_{a=1}^{p} \beta_a x_{ait} + \beta_0 + c_i + u_{it}$$

ここに、$i = 1, \cdots, n$ は個体の識別記号、$t = 1, \cdots, T$ は時間の識別番号である。通常の回帰分析と異なるところは、時間について変化しない固定効果 c_i を含むことである。但し、固定効果と誤差項については以下の仮定を置き、固定効果については直接には観測されないものとする。

$$\bar{c} = 0, \ \bar{u}_i = \bar{u} = 0$$

最初の式から観測されない c_i を消すために、次の変換を実行し、

$$\tilde{y}_{it} = y_{it} - \bar{y}_i, \quad \tilde{x}_{ait} = x_{ait} - \bar{x}_{ai}, \quad \tilde{u}_{it} = u_{it} - \bar{u}_i = u_{it}$$

以下の関係を得る。

$$\tilde{y}_{it} = \sum_{a=1}^{p} \beta_a \tilde{x}_{ait} + u_{it}$$

この関係を使って回帰分析を実行し、回帰係数 $\hat{\beta}_a$ を推定する。最後に、推定された回帰係数を使って、定数項、固定効果について、以下のように求める。

$$\hat{\beta}_0 = \bar{y} - \sum_{a=1}^{p} \hat{\beta}_a \bar{x}_a, \qquad \hat{c}_i = \bar{y}_i - \sum_{a=1}^{p} \hat{\beta}_a \bar{x}_{ai} - \hat{\beta}_0$$

定義や式の詳細については、最後の節にまとめておく。

2.1 パネルデータ分析の理論

パネルデータの目的変数を y_{it} （$i = 1, \cdots, n$， $t = 1, \cdots, T$ ）、説明変数を x_{ait} （$a = 1, \cdots, p$）、個体の固定効果を c_i 、時間効果を d_t 、定数項を β_0 、誤差を u_{it} として、以下のモデルを考える。

$$Y_{it} = \sum_{a=1}^{p} \beta_a x_{ait} + \beta_0 + \delta_f c_i + \delta_t d_t$$

$$y_{it} = Y_{it} + u_{it} = \sum_{a=1}^{p} \beta_a x_{ait} + \beta_0 + \delta_f c_i + \delta_t d_t + u_{it}$$

ここに、δ_f は固定効果を考える場合は 1、考えない場合は 0 を与える定数で、δ_t は時間効果を考える場合は 1、考えない場合は 0 を与える定数である。また、固定効果、時間効果、誤差について、以下を仮定する。

$$\bar{c} = \bar{d} = 0, \ \bar{u}_i = \bar{u}_t = \bar{u} = 0$$

また､変数については以下のような変換を考える(固定効果がある場合は時間平均を引き､時間効果がある場合は個体平均を引く)。

$$\tilde{y}_{it} = y_{it} - \delta_f \bar{y}_i - \delta_t \bar{y}_t + \delta_f \delta_t \bar{y}, \quad \tilde{x}_{ait} = x_{ait} - \delta_f \bar{x}_{ai} - \delta_t \bar{x}_{at} + \delta_f \delta_t \bar{x}_a,$$

$$\tilde{u}_{it} = u_{it} - \delta_f \bar{u}_i - \delta_t \bar{u}_t + \delta_f \delta_t \bar{u} = u_{it}$$

ここに、

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T}\left[\sum_{a=1}^{p}\beta_a x_{ait} + \delta_f c_i + \delta_t d_t + \beta_0 + u_{it}\right] = \sum_{a=1}^{p}\beta_a \bar{x}_{ai} + \delta_f c_i + \beta_0$$

$$\bar{y}_t = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{a=1}^{p}\beta_a x_{ait} + \delta_f c_i + \delta_t d_t + \beta_0 + u_{it}\right] = \sum_{a=1}^{p}\beta_a \bar{x}_{at} + \delta_t d_t + \beta_0$$

$$\bar{y} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left[\sum_{a=1}^{p}\beta_a x_{ait} + \delta_f c_i + \delta_t d_t + \beta_0 + u_{it}\right] = \sum_{a=1}^{p}\beta_a \bar{x}_a + \beta_0$$

この変換によって、固定効果と時間効果の項は消え、以下のような関係を得る。

$$\tilde{y}_{it} = \sum_{a=1}^{p}\beta_a \tilde{x}_{ait} + \delta_f c_i - \delta_f^2 c_i + \delta_t d_t - \delta_t^2 d_t + \beta_0(1-\delta_f - \delta_t + \delta_f \delta_t) + u_{it}$$

$$= \sum_{a=1}^{p}\beta_a \tilde{x}_{ait} + (1-\delta_f)(1-\delta_t)\beta_0 + u_{it}$$

これにより、予測値は以下のようになる。

$$\tilde{Y}_{it} = \sum_{a=1}^{p}\hat{\beta}_a \tilde{x}_{ait} + (1-\delta_f)(1-\delta_t)\hat{\beta}_0$$

ここで実際のプログラムでは、$(1-\delta_f)(1-\delta_t)\hat{\beta}_0$ 部分を改めて $\hat{\beta}_0$ として計算している。後に述べるように､固定効果や時間効果がある場合､最小 2 乗法の計算から予測値 $\hat{\beta}_0$ は自動的に 0 になる。

実測値と予測値の差の 2 乗を以下のように L とする。

$$L = \sum_{i=1}^{n}\sum_{t=1}^{T}\left(\tilde{y}_{it} - \tilde{Y}_{it}\right)^2$$

これを最小化することから、

$$\frac{\partial L}{\partial \beta_0} = -2\sum_{i=1}^{n}\sum_{t=1}^{T}\left(\tilde{y}_{it} - \sum_{b=1}^{p}\hat{\beta}_b \tilde{x}_{bit} - \hat{\beta}_0\right) = 0$$

$$\frac{\partial L}{\partial \beta_a} = -2\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{x}_{ait}\left(\tilde{y}_{it} - \sum_{b=1}^{p}\hat{\beta}_b \tilde{x}_{bit} - \hat{\beta}_0\right) = 0$$

ここで、固定効果や時間効果がある場合、

$$\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{y}_{it}=0,\quad \sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{x}_{bit}=0$$

であることから、第 1 式より自動的に $\hat{\beta}_0=0$ が示される。

最小化の方程式は、まとめて、

$$\tilde{\mathbf{y}}=\begin{pmatrix}\tilde{y}_{11}\\ \tilde{y}_{12}\\ \vdots\\ \tilde{y}_{nT}\end{pmatrix},\quad \tilde{\mathbf{X}}=\begin{pmatrix}1 & \tilde{x}_{111} & \cdots & \tilde{x}_{p11}\\ 1 & \tilde{x}_{112} & \cdots & \tilde{x}_{p12}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & \tilde{x}_{1nT} & \cdots & \tilde{x}_{pnT}\end{pmatrix},\quad \mathbf{u}=\begin{pmatrix}u_{11}\\ u_{12}\\ \vdots\\ u_{nT}\end{pmatrix}$$

のように定義すると、以下のように書かれる。

$$\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}=\tilde{\mathbf{X}}'\tilde{\mathbf{y}}$$

これを解いて、以下を得る。

$$\hat{\boldsymbol{\beta}}=(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}}$$

この計算過程から、残差 $\hat{u}_{it}$ に対する以下の制約が得られる。

$$\sum_{i=1}^{n}\sum_{t=1}^{T}\hat{u}_{it}=0,\qquad \sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{x}_{ait}\hat{u}_{it}=0$$

これは、重回帰分析における誤差項の性質である。

固定効果や時間効果がある場合、以下の回帰式から、

$$\tilde{y}_{it}=\sum_{a=1}^{p}\hat{\beta}_a\tilde{x}_{ait}+\hat{\beta}_0+\hat{u}_{it}$$

次の制約も追加される。

固定効果から $\bar{\hat{u}}_i=\dfrac{1}{T}\sum_{t=1}^{T}\hat{u}_{it}=0$，　時間効果から $\bar{\hat{u}}_t=\dfrac{1}{n}\sum_{i=1}^{n}\hat{u}_{it}=0$

すべての仮定を考えると、残差 $\hat{u}_{it}$ の自由度は以下となる。

$$D=nT-p-\delta_f(n-1)-\delta_t(T-1)-1$$

次に、パラメータの標準誤差について考える。回帰式 $\tilde{\mathbf{y}}=\tilde{\mathbf{X}}\boldsymbol{\beta}+\mathbf{u}$ より、

$$\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}=(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{u}$$

これらを成分で表示すると以下となるが、

$$(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})_{ab}=\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{x}_{ait}\tilde{x}_{bit},\quad (\tilde{\mathbf{X}}'\mathbf{u})_a=\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{x}_{ait}u_{it}$$

$\tilde{\mathbf{w}}=\tilde{\mathbf{X}}'\mathbf{u}$ とすると、平均と共分散は以下となる。

$$E\left[\tilde{w}_a\right]=E\left[(\tilde{\mathbf{X}}'\mathbf{u})_a\right]=\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{x}_{ait}E\left[u_{it}\right]=0$$

$$Cov\left[\tilde{w}_a,\tilde{w}_b\right]=Cov\left[(\tilde{\mathbf{X}}'\mathbf{u})_a,(\tilde{\mathbf{X}}'\mathbf{u})_b\right]=E\left[\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{x}_{ait}u_{it}\sum_{j=1}^{n}\sum_{t'=1}^{T}\tilde{x}_{bjt'}u_{jt'}\right]$$

$$=E\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T}\sum_{t'=1}^{T}\tilde{x}_{ait}u_{it}\tilde{x}_{bjt'}u_{jt'}\right]$$

個体に対する独立性を仮定した場合（時間に関しては仮定しない）、

$$E\left[u_{it}u_{jt'}\right]=E\left[u_{it}u_{it'}\right]\delta_{ij},\quad \tilde{v}_{ait}=\tilde{x}_{ait}u_{it},\quad \eta_{ai}=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\tilde{v}_{ait}$$ とすると

$$Cov\left[\tilde{w}_a,\tilde{w}_b\right]=E\left[\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{t'=1}^{T}\tilde{x}_{ait}u_{it}\tilde{x}_{bit'}u_{it'}\right]=E\left[\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{v}_{ait}\sum_{t'=1}^{T}\tilde{v}_{ait'}\right]$$

$$=TE\left[\sum_{i=1}^{n}\eta_{ai}\eta_{bi}\right]=nT\sigma_{\eta_a\eta_b}$$

$$\rightarrow\frac{nT}{n-1}\sum_{i=1}^{n}\hat{\eta}_{ai}\hat{\eta}_{bi}=\frac{n}{n-1}\sum_{i=1}^{n}\left(\sum_{t=1}^{T}\tilde{v}_{ait}\right)\left(\sum_{t'=1}^{T}\tilde{v}_{ait'}\right)$$

時間に関する独立性を仮定した場合（個体に関しては仮定しない）

$$E\left[u_{it}u_{jt'}\right]=E\left[u_{it}u_{jt}\right]\delta_{tt'},\quad \tilde{v}_{ait}=\tilde{x}_{ait}u_{it},\quad \gamma_{at}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{v}_{ait}$$ とすると

$$Cov\left[\tilde{w}_a,\tilde{w}_b\right]=E\left[\sum_{t=1}^{T}\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{x}_{ait}u_{it}\tilde{x}_{bjt}u_{jt}\right]=E\left[\sum_{t=1}^{T}\sum_{i=1}^{n}\tilde{v}_{ait}\sum_{j=1}^{n}\tilde{v}_{bjt}\right]$$

$$=nE\left[\sum_{t=1}^{T}\gamma_{at}\gamma_{bt}\right]=nT\sigma_{\gamma_a\gamma_b}$$

$$\rightarrow\frac{nT}{T-1}\sum_{t=1}^{T}\hat{\gamma}_{at}\hat{\gamma}_{bt}=\frac{T}{T-1}\sum_{t=1}^{T}\left(\sum_{i=1}^{n}\tilde{v}_{ait}\right)\left(\sum_{j=1}^{n}\tilde{v}_{ajt}\right)$$

個体と時間に関する独立性を仮定した場合（通常の不均一分散）

$$E\left[u_{it}u_{jt'}\right]=E\left[u_{it}u_{it}\right]\delta_{ij}\delta_{tt'},\quad \tilde{v}_{ait}=\tilde{x}_{ait}u_{it}$$ とすると

$$Cov\left[\tilde{w}_a,\tilde{w}_b\right]=E\left[\sum_{t=1}^{T}\sum_{i=1}^{n}\tilde{x}_{ait}u_{it}\tilde{x}_{bit}u_{it}\right]=E\left[\sum_{t=1}^{T}\sum_{i=1}^{n}\tilde{v}_{ait}\tilde{v}_{bit}\right]=nT\sigma_{\tilde{v}_a\tilde{v}_b}$$

$$\to\frac{nT}{nT-p-1}\sum_{t=1}^{T}\sum_{i=1}^{n}\tilde{v}_{ait}\tilde{v}_{bit}$$

個体も時間も独立性を仮定しない場合

$$\tilde{v}_{ait}=\tilde{x}_{ait}u_{it}\,,\quad \phi_{at}=\frac{1}{\sqrt{nT}}\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{v}_{ait}\qquad \text{とすると}$$

$$Cov\left[\tilde{w}_a,\tilde{w}_b\right]=E\left[\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{x}_{ait}u_{it}\sum_{j=1}^{n}\sum_{t'=1}^{T}\tilde{x}_{bjt'}u_{jt'}\right]=nT\sigma_{\phi_a\phi_b}$$

$$\to nT\hat{\phi}_a\hat{\phi}_b=\sum_{i=1}^{n}\sum_{t=1}^{T}\hat{\tilde{v}}_{ait}\sum_{j=1}^{n}\sum_{t'=1}^{T}\hat{\tilde{v}}_{bjt'}$$

上で述べたことを利用して、改めてパラメータ $\hat{\boldsymbol{\beta}}$ の共分散を求める。

$$\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}=(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{u}$$

より、$\mathbf{G}=(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}$ とおくと $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}=\mathbf{G}\tilde{\mathbf{w}}$

これから以下の関係を得る。

$$Cov\left[\hat{\beta}_a,\hat{\beta}_b\right]=Cov\left[(\mathbf{G}\tilde{\mathbf{w}})_a,(\mathbf{G}\tilde{\mathbf{w}})_b\right]=E\left[\sum_{c=1}^{p}g_{ac}\tilde{w}_c\sum_{d=1}^{p}g_{bd}\tilde{w}_d\right]$$

$$=\sum_{c=1}^{n}\sum_{d=1}^{n}g_{ac}g_{bd}Cov\left[\tilde{w}_c,\tilde{w}_d\right]$$

個体に対する独立性を仮定した場合（時間に関しては仮定しない）

$$Cov\left[\hat{\beta}_a,\hat{\beta}_b\right]\to\frac{nT}{n-1}\sum_{c=1}^{n}\sum_{d=1}^{n}g_{ac}g_{bd}\sum_{i=1}^{n}\hat{\eta}_{ci}\hat{\eta}_{di}$$

$$=\frac{n}{n-1}\sum_{c=1}^{n}\sum_{d=1}^{n}g_{ac}g_{bd}\sum_{i=1}^{n}\left(\sum_{t=1}^{T}\hat{\tilde{v}}_{cit}\right)\left(\sum_{t'=1}^{T}\tilde{v}_{dit'}\right)$$

他も同様であるのでここでは省略する。

ここで、$p=1$の固定効果モデルで、個体に対する独立性を仮定すると

$$\tilde{\mathbf{X}}'\tilde{\mathbf{X}}=(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})_{11}=\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{x}_{1it}^2\ \text{より、}\ \mathbf{G}=g=\frac{1}{(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})_{11}}=\frac{1}{\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{x}_{1it}^2}=\frac{1}{nTQ_{\tilde{x}}}$$

$$Var\left[\hat{\beta}\right]=Var\left[g\tilde{w}_1\right]=\frac{1}{n^2T^2Q_{\tilde{x}}^2}Var\left[\tilde{w}_1\right]$$

$$\to\frac{nT}{n-1}\frac{1}{n^2T^2\hat{Q}_{\tilde{x}}^2}\sum_{i=1}^{n}\hat{\eta}_{1i}^2=\frac{n}{n-1}\frac{1}{n^2T^2\hat{Q}_{\tilde{x}}^2}\sum_{i=1}^{n}\left(\sum_{t=1}^{T}\hat{\tilde{v}}_{1it}\right)^2$$

有効性の検定には以下の関係を用いるが、

$$F=(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}^{-1}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\sim F_{k,\infty}$$

個体と時間に関する独立性を仮定しない方法で作られた共分散行列は正則ではなく、有効性の検定には利用できない。

固定効果と時間効果の求め方

固定効果や時間効果がある場合、その値を求める必要がある。回帰式は以下である。

$$y_{it}=\sum_{a=1}^{p}\beta_a x_{ait}+c_i+d_t+\beta_0+u_{it}$$

回帰分析の結果は以下の式になる。

$$\tilde{y}_{it}=\sum_{a=1}^{p}\hat{\beta}_a\tilde{x}_{ait}+\hat{u}_{it}$$

左辺と右辺第 1 項の関係から、

$$\bar{\hat{u}}_i=\bar{\hat{u}}_t=\bar{\hat{u}}=0$$

この関係を元の式に用いると

$$y_{it}=\sum_{a=1}^{p}\hat{\beta}_a x_{ait}+\hat{\beta}_0+c_i+d_t+\hat{u}_{it}$$

これより

$$\bar{y}=\sum_{a=1}^{p}\hat{\beta}_a\bar{x}_a+\hat{\beta}_0\quad\to\quad\hat{\beta}_0=\bar{y}-\sum_{a=1}^{p}\hat{\beta}_a\bar{x}_a$$

$$\bar{y}_i=\sum_{a=1}^{p}\hat{\beta}_a\bar{x}_{ai}+\hat{\beta}_0+\hat{c}_i\quad\to\quad\hat{c}_i=\bar{y}_i-\sum_{a=1}^{p}\hat{\beta}_a\bar{x}_{ai}-\hat{\beta}_0$$

$$\bar{y}_t=\sum_{a=1}^{p}\hat{\beta}_a\bar{x}_{at}+\hat{\beta}_0+\hat{d}_t\quad\to\quad\hat{d}_t=\bar{y}_t-\sum_{a=1}^{p}\hat{\beta}_a\bar{x}_{at}-\hat{\beta}_0$$

第 1 式の関係を用いると、$\bar{\hat{c}}=\bar{\hat{d}}=0$を得る。

以上述べた処理は欠損値がある場合、欠損データを除いた平均を取ることによって同様に実行できるが、その平均に偏りがある場合、十分な注意が必要である。例えば時間的に増加している変数の時間的に最初や最後のデータの欠損は平均にずれを生じるため、そのまま平均を取ることには問題がある。その個体のデータ全体を取り除く必要があるだろう。

2.2　プログラムの利用法

メニュー［分析－多変量解析他－経済・経営手法－パネルデータ分析］を選択すると、図 1 のような分析実行画面が表示される。

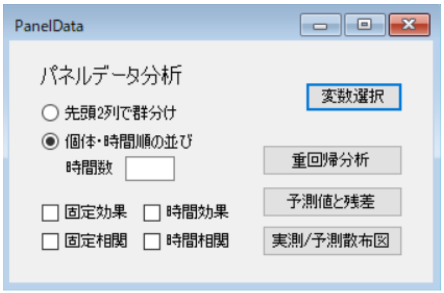

図 1　分析実行画面

データは通常の重回帰分析と同じであるが、レコード配置は、「個体・時間順の並び」の場合、図 2 のように、個体 1・年次 1, 個体 1・年次 2, … , 個体 n・年次 1, …, 個体 n・年次 T の順番で目的変数, 説明変数 1 , 説明変数 2 , … が順番に並び、「先頭 2 列で群分け」の場合、図 3 のように、個体分類, 時間分類, 目的変数, 説明変数 1 , 説明変数 2 , … のようにデータが並ぶ。

	目的変数	説明変数1	説明変数2	説明変数3
1	127	650	18	18
2	153	800	15	18
3	183	900	20	23
4	177	1000	22	14
5	83	400	16	22
6	106	480	20	25
7	135	550	18	19
8	147	680	18	10
9	100	500	15	19
10	125	560	23	15

図 2　「個体・時間順の並び」のデータ形式（パネルデータ分析 1.txt）

	個体	年度	目的変数	説明変数1	説明変数2	説明変数3
1	1	2000	127	650	18	18
2	1	2005	153	800	15	18
3	1	2010	183	900	20	23
4	1	2015	177	1000	22	14
5	2	2000	83	400	16	22
6	2	2005	106	480	20	25
7	2	2010	135	550	18	19
8	2	2015	147	680	18	10
9	3	2000	100	500	15	19
10	3	2005	125	560	23	15
11	3	2010	137	650	23	22

図 3　「先頭 2 列で群分け」のデータ形式（パネルデータ分析 1.txt）

図 2 の形式では、データを欠損値なく、順序通りに並べる必要があるが、図 3 の形式では、データの順序が変わっていても、欠損値があっても対応可能である。ただ、時間平均や個体平均を取る場合、欠損値を除いた個数で平均をとるため、偏りがあるような欠損値の場合（例えば、時間と共に増大するデータの最初の値や最後の値の欠損など）、結果の正当性に問題が出る恐れもある。

図 2 の場合、変数選択ですべての列を選び、「時間数」を 4 と入力して、その他の設定は何もせず、「重回帰分析」ボタンをクリックした結果を図 4 に示す。

偏回帰係数と検定

目的変数	偏回帰係数	標準化係数	標準誤差	z統計量	確率値	95%下限	95%上限
説明変数1	0.1410	0.8489	0.0130	10.8450	0.0000	0.1155	0.1665
説明変数2	2.4932	0.2642	0.6087	4.0956	0.0000	1.3001	3.6863
説明変数3	-0.2353	-0.0436	0.2818	-0.8349	0.4038	-0.7876	0.3170
切片	-0.0319	0.0000	12.0772	-0.0026	0.9979	-23.7028	23.6391
R	0.957	R^2	0.916				

図 4 効果も相関も考えない場合の結果

これは、通常の重回帰分析の結果に一致している。

次に、時間に関してデータ間に相関がある（独立でない）場合について分析を実行する。「時間相関」にチェックを入れて、「重回帰分析」ボタンをクリックすると図 5 に示す結果となる。

偏回帰係数と検定

目的変数	偏回帰係数	標準化係数	標準誤差	z統計量	確率値	95%下限	95%上限
説明変数1	0.1410	0.8489	0.0064	22.0403	0.0000	0.1285	0.1535
説明変数2	2.4932	0.2642	0.5638	4.4218	0.0000	1.3881	3.5983
説明変数3	-0.2353	-0.0436	0.3494	-0.6735	0.5007	-0.9200	0.4495
切片	-0.0319	0.0000	11.5072	-0.0028	0.9978	-22.5855	22.5218
R	0.957	R^2	0.916				

図 5 時間相関がある場合の結果

次に、固定効果と時間相関がある場合の例を示す。「固定効果」チェックボックスと「時間相関」チェックボックスにチェックを入れて「重回帰分析」ボタンをクリックすると、図 6 のような結果となる。

偏回帰係数と検定

目的変数	偏回帰係数	標準化係数	標準誤差	z統計量	確率値	95%下限	95%上限
説明変数1	0.1505	0.9061	0.0174	8.6690	0.0000	0.1165	0.1845
説明変数2	2.6666	0.2826	0.7125	3.7425	0.0002	1.2701	4.0632
説明変数3	-0.1792	-0.0332	0.3377	-0.5308	0.5955	-0.8410	0.4826
切片	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000
R	0.962	R^2	0.925				

図 6 固定効果と時間相関がある場合の結果

固定効果と時間相関がある場合の、予測値について示す。「予測値と残差」ボタンをクリックすると、図 7 のような結果を得る。

予測値と残差

目的変数	実測値	部分予測値	固定効果	時間効果	定数	予測値	残差
1	127	142.5975	-1.8880	0.0000	-10.8833	129.8262	-2.8262
2	153	157.1725	-1.8880	0.0000	-10.8833	144.4012	8.5988
3	183	184.6594	-1.8880	0.0000	-10.8833	171.8881	11.1119
4	177	206.6558	-1.8880	0.0000	-10.8833	193.8844	-16.8844
5	83	98.9225	4.6512	0.0000	-10.8833	92.6904	-9.6904
6	106	121.0912	4.6512	0.0000	-10.8833	114.8591	-8.8591
7	135	127.3684	4.6512	0.0000	-10.8833	121.1362	13.8638
8	147	148.5464	4.6512	0.0000	-10.8833	142.3143	4.6857
9	100	111.8435	-1.7283	0.0000	-10.8833	99.2319	0.7681
10	125	142.9235	-1.7283	0.0000	-10.8833	130.3119	-5.3119

図 7 固定効果と時間相関がある場合の予測値と残差

図 7 により、本章 1 節の最初に述べた回帰式の構造がよく分かる。部分予測値は、個体効果や時間効果を除いた予測値で、それにこれらの効果や定数を含めると実際の予測値となる。これは最後から 2 列目に表示されている。

最後に、「実測・予測散布図」ボタンをクリックして実測値と予測値をグラフに描くと図 8 のようになる。

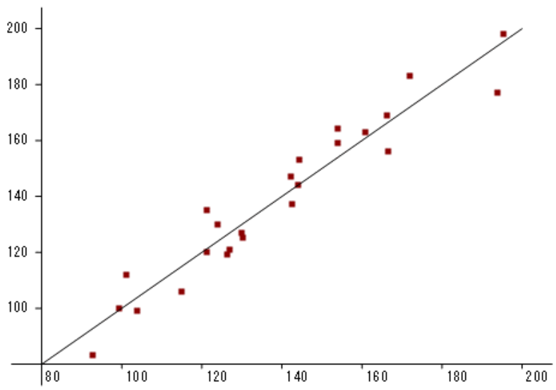

図 8 実測値/予測値グラフ

3. 操作変数回帰分析

操作変数回帰分析は観測されない除外されたバイアス要因から影響を受ける説明変数から、その影響を取り除く手法である。今、目的変数が２つの説明変数で表される重回帰分析を考える。２つの説明変数のうち１つは目的変数に直接影響を及ぼし、他からの影響は考えられないものとする。しかし、もう１つは重回帰式に取り入れられないある要因から影響を受け、同時に目的変数もその要因から影響を受けるものとする。この要因により、目的変数とこの説明変数の間には直接的な影響の他に擬似相関が発生する。我々は前者の直接的な影響を与える説明変数を外生変数と呼び、後者のある要因から影響を受ける説明変数を内生変数と呼ぶ。また、この内生変数に影響を及ぼす要因を欠落バイアス要因（変数）と呼ぶ。図 1 にこれらの関係を示す。

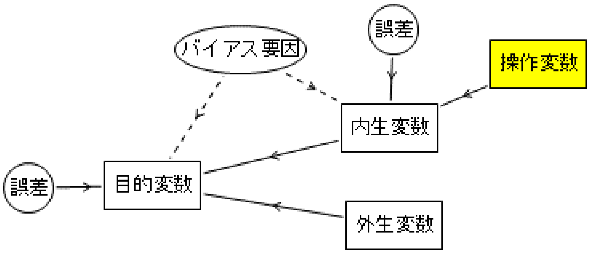

図１　操作変数回帰分析の関係

このようなバイアスが存在する場合、２つの説明変数による重回帰分析は正しい回帰係数（直接影響）を与えない。本来重回帰分析では誤差と説明変数の間に相関がないように回帰係数を推定するからである。また、この潜在変数を考えるようなモデルは共分散構造分析の対象のように思われるかも知れないが、これは解の識別が不可能なモデルで、安定した解を求めることはできない。この状況を解決する方法が操作変数回帰分析である。

操作変数回帰分析では内生変数だけに影響を与え、目的変数に直接影響を与えないような観測可能な変数を考える。この変数を操作変数と呼ぶ。分析方法は、まず内生変数を操作変数と他の外生変数で回帰し、その予測値を内生変数の実測値の代わりにして重回帰分析を行うというものである。もちろん内生変数や操作変数は複数あっても構わない。ただ、操作変数の数は内生変数の数以上である必要がある。

ここでは操作変数回帰分析の一般的な手順を数式的に簡単に示しておこう。操作変数回帰分析では目的変数を以下のような線形の式で予測する。ここに、変数 $x_{i\lambda}$ $(i=1,\cdots,k)$ は誤差項 ε_λ と相関する内生変数と呼ばれる変数で、変数 $w_{i\lambda}$ $(i=1,\cdots,r)$ は誤差項と相関しない外生変数と呼ばれる変数である。

$$y_\lambda = \sum_{i=1}^{k} \beta_i x_{i\lambda} + \sum_{i=1}^{r} \beta_{k+i} w_{i\lambda} + \beta_0 + \varepsilon_\lambda \quad (1)$$

一般の回帰分析では、変数はすべて外生変数である。

操作変数回帰分析では一般に２段階法という手法が使われる。すなわち、まず各内生変数 $x_{i\lambda}$ を操作変数と呼ばれる、誤差項 ε_λ と無相関な変数 $z_{i\lambda}$ $(i=1,\cdots,m \geq k)$ と外生変数 $w_{i\lambda}$ で予測する。

$$x_{i\lambda} = \sum_{j=1}^{m} c_j^{(i)} z_{j\lambda} + \sum_{j=1}^{r} c_{r+j}^{(i)} w_{j\lambda} + c_0^{(i)} + v_{i\lambda} = \hat{x}_{i\lambda} + v_{i\lambda}$$

ここで、$\hat{x}_{i\lambda}$ は回帰の予測値である。次にこの予測値で(1)式の内生変数を置き換え、回帰分析を実行する。

$$y_\lambda = \sum_{i=1}^{k} b_i \hat{x}_{i\lambda} + \sum_{i=1}^{r} b_{k+i} w_{i\lambda} + b_0 + u_\lambda$$

この推定されたパラメータ b_i, b_{k+i}, b_0 は $N \to \infty$ で、真の値 $\beta_i, \beta_{k+i}, \beta_0$ に一致することが知られている。ここで、パラメータの区間推定などに利用される誤差項 e_λ については、以下の式から求められる。

$$y_\lambda = \sum_{i=1}^{k} b_i x_{i\lambda} + \sum_{i=1}^{r} b_{k+i} w_{i\lambda} + b_0 + e_\lambda$$

ここでは得られたパラメータの値はそのままで、内生変数の予測値を元の実測値に変えている。これは、各パラメータが $N \longrightarrow \infty$ で真の値に近づくことから、誤差 e_λ も真の誤差 ε_λ に近づくことによる。これ以上の詳細は後の理論的解説のところで紹介する。

3.1　操作変数回帰分析の理論

欠落バイアスの影響により、回帰式の説明変数と誤差項との間に相関があるとき、通常の回帰分析では回帰係数の推定値に問題がある。しかし、一部の説明変数に影響を与え、誤差項と無相関な変数を見つけられれば、その変数を利用して、正しい回帰係数を求めることができる。これが操作変数回帰分析である。説明変数の中で誤差項と無相関な変数を外生変数、誤差項と相関がある変数を内生変数と呼ぶ。内生変数だけに影響を与え、誤差項と無相関な変数を操作変数と呼ぶ。ここでは目的変数を y_λ ($\lambda = 1, \cdots, N$)、内生変数を $x_{i\lambda}$ $(i = 1, \cdots, k)$ 、外生変数を $w_{i\lambda}$ $(i = 1, \cdots, r)$ 、操作変数を $z_{i\lambda}$ $(i = 1, \cdots, m)$ とし、これらの変数を以下のように行列表示する。ここで個体間の相関はないものと仮定する。

$$\mathbf{y}(N \times 1) = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

$$\mathbf{X}(N \times (k + r + 1)) = \begin{pmatrix} 1 & x_{11} & \cdots & x_{k1} & w_{11} & \cdots & w_{r1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1N} & \cdots & x_{kN} & w_{1N} & \cdots & w_{rN} \end{pmatrix},$$

$$\mathbf{Z}(N \times (m + r + 1)) = \begin{pmatrix} 1 & z_{11} & \cdots & z_{m1} & w_{11} & \cdots & w_{r1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_{1N} & \cdots & z_{mN} & w_{1N} & \cdots & w_{rN} \end{pmatrix}$$

真の回帰係数と誤差を $\boldsymbol{\beta}((k + r + 1) \times 1)$, $\boldsymbol{\varepsilon}(N \times 1)$ とすると、回帰式は以下のように書ける。

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

この式を用いてそのまま最小２乗法を実行すると、$\boldsymbol{\varepsilon}$ の推定値 $\hat{\boldsymbol{\varepsilon}}$ と $\mathbf{X}$ は結果として独立となり、真の回帰係数 $\boldsymbol{\beta}$ の推定値 $\hat{\boldsymbol{\beta}}$ は得られない。これ以降ある量の真の値を a とすると

き、その推定値を $\hat{a}$ と書くことにする。

この問題を解決するために、操作変数回帰分析の２段階法では以下のようなことを考える。まず、第１段階として、説明変数の中の内生変数について、すべての操作変数と外生変数で回帰する。これを行列表示すると以下となる。

$$\mathbf{X} = \mathbf{ZC} + \mathbf{V} \tag{2}$$

ここで $\mathbf{C}$ は係数行列、$\mathbf{V}$ は誤差行列である。但し、$\mathbf{X}$ に含まれる定数と外生変数については恒等式とするので、$\mathbf{C}$ と $\mathbf{V}$ については以下のようにする。

$$\mathbf{C}((m+r+1)\times(k+r+1)) = \begin{pmatrix} 1 & \mathbf{c_0}(1\times k) & \mathbf{0}(1\times r) \\ \mathbf{0}(m\times 1) & \mathbf{c_z}(m\times k) & \mathbf{0}(m\times r) \\ \mathbf{0}(r\times 1) & \mathbf{c_w}(r\times k) & \mathbf{I}_r \end{pmatrix}$$

$$\mathbf{V}(N\times(k+r+1)) = \begin{pmatrix} 0 & v_{11} & \cdots & v_{k1} & 0 & \cdots & 0 \\ 0 & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & v_{1N} & \cdots & v_{kN} & 0 & \cdots & 0 \end{pmatrix}$$

これらの推定値を $\hat{\mathbf{C}}$ と $\hat{\mathbf{V}}$ とすると、推定結果は以下のように書ける。

$$\mathbf{X} = \mathbf{Z}\hat{\mathbf{C}} + \hat{\mathbf{V}} = \hat{\mathbf{X}} + \hat{\mathbf{V}}$$

次に第２段階として、予測値 $\hat{\mathbf{X}}$ を使って以下のような回帰式を考える。

$$\mathbf{y} = \hat{\mathbf{X}}\mathbf{b} + \mathbf{u} \tag{3}$$

推定値の関係は以下となる。

$$\mathbf{y} = \hat{\mathbf{X}}\hat{\mathbf{b}} + \hat{\mathbf{u}}$$

この推定値 $\hat{\mathbf{b}}$ は $N \to \infty$ で真の回帰係数 $\boldsymbol{\beta}$ に収束することが知られている。これを証明する。

(2)式の回帰係数の推定値については、

$$\hat{\mathbf{C}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$$

と書けることから、$\mathbf{P_Z} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ として以下を得る。

$$\hat{\mathbf{X}} = \mathbf{Z}\hat{\mathbf{C}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} = \mathbf{P_Z}\mathbf{X}, \qquad \hat{\mathbf{X}}'\hat{\mathbf{X}} = \mathbf{X}'\mathbf{P_Z}\mathbf{X}$$

ここに、ある行列 $\mathbf{A}$ に対して、$\mathbf{P_A} \equiv \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ と定義する。この $\mathbf{P_A}$ はべき等行列（$\mathbf{P_A}\mathbf{P_A} = \mathbf{P_A}$）である。この関係を使うと $\hat{\mathbf{b}}$ は以下のように書ける。

$$\begin{aligned}\hat{\mathbf{b}} &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y} = (\mathbf{X}'\mathbf{P_Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P_Z}\mathbf{y} \\ &= (\mathbf{X}'\mathbf{P_Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P_Z}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{P_Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P_Z}\boldsymbol{\varepsilon}\end{aligned}$$

上の関係より、【公式 3】を使うと

$$\hat{\mathbf{b}} \sim N(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{P_Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P_Z}\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\mathbf{P_Z}\mathbf{X}(\mathbf{X}'\mathbf{P_Z}\mathbf{X})^{-1})$$

ここで、外生変数は誤差項 $\boldsymbol{\varepsilon}$ と無相関であるという仮定から、

$$\mathbf{Z}'\boldsymbol{\varepsilon} \xrightarrow[N\to\infty]{} \mathbf{0}\text{，即ち、}\mathbf{P_Z}\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\mathbf{P_Z} \xrightarrow[N\to\infty]{} \mathbf{0}$$

よって 2 段階法で求めた回帰係数は $N\to\infty$ で真の回帰係数に一致する。

$$\hat{\mathbf{b}} \xrightarrow[N\to\infty]{} \boldsymbol{\beta}$$

この関係を使うと以下となり、

$$\hat{\mathbf{e}} \equiv \mathbf{y}-\mathbf{X}\hat{\mathbf{b}} \xrightarrow[N\to\infty]{} \mathbf{y}-\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\varepsilon}$$

この $\hat{\mathbf{e}}$ を用いて $\boldsymbol{\varepsilon}$ の推定値とできることが分かる。ここで注意することは、$\hat{\mathbf{e}}$ の定義では $\hat{\mathbf{X}}$ ではなく、$\mathbf{X}$ の値をそのまま使うことである。以上より、

$$\begin{aligned}\hat{\mathbf{b}} &\sim N(\boldsymbol{\beta},(\mathbf{X}'\mathbf{P_Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P_Z}\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\mathbf{P_Z}\mathbf{X}(\mathbf{X}'\mathbf{P_Z}\mathbf{X})^{-1})\\ &\to N(\boldsymbol{\beta},(\mathbf{X}'\mathbf{P_Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P_Z}\hat{\mathbf{e}}\hat{\mathbf{e}}'\mathbf{P_Z}\mathbf{X}(\mathbf{X}'\mathbf{P_Z}\mathbf{X})^{-1})\end{aligned}$$

有効性の検定では、重回帰分析で利用した以下の式を用いる。

$$F = (\mathbf{R}\hat{\mathbf{b}}-\mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{P_Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P_Z}\hat{\mathbf{e}}\hat{\mathbf{e}}'\mathbf{P_Z}\mathbf{X}(\mathbf{X}'\mathbf{P_Z}\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\mathbf{b}}-\mathbf{r})\big/q \sim F_{q,\infty}$$

ここに、

$$\mathbf{R}((k+r)\times(k+r+1)) = \left[\mathbf{0}((k+r)\times 1)\quad \mathbf{I}((k+r)\times(k+r))\right],$$

$$\mathbf{r}((k+r)\times 1) = \mathbf{0}$$

プログラムには組み込んでいないが、これは $\mathbf{R}$ と $\mathbf{r}$ の取り方によって、一般的な結合仮説の検定にも利用できる。

以下の関係を使うと、

$$\begin{aligned}\hat{\mathbf{b}} &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})\\ &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'((\hat{\mathbf{X}}+\hat{\mathbf{V}})\boldsymbol{\beta}+\boldsymbol{\varepsilon}) = \boldsymbol{\beta}+(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\boldsymbol{\varepsilon}\end{aligned}$$

$\hat{\mathbf{b}}$ の分布は以下のように書くこともできる。

$$\hat{\mathbf{b}} \sim N(\boldsymbol{\beta},(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\hat{\mathbf{X}}(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}) \to N(\boldsymbol{\beta},(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\hat{\mathbf{e}}\hat{\mathbf{e}}'\hat{\mathbf{X}}(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1})$$

我々のプログラムではこの関係を利用している。

余談であるが、説明変数、操作変数共に 1 つの場合、

$$\mathbf{P_Z} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' = \mathbf{Z}\mathbf{Z}'/\mathbf{Z}'\mathbf{Z}$$

より、以下となる。

$$\begin{aligned}\hat{\mathbf{b}} &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y} = [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}\\ &= \mathbf{Z}'\mathbf{y}/\mathbf{X}'\mathbf{Z} = \boldsymbol{\beta}+\mathbf{Z}'\boldsymbol{\varepsilon}/\mathbf{X}'\mathbf{Z} \to \boldsymbol{\beta}+\mathbf{Z}'\hat{\mathbf{e}}/\mathbf{X}'\mathbf{Z} \sim N(\boldsymbol{\beta},(\mathbf{Z}'\hat{\mathbf{e}}/\mathbf{X}'\mathbf{Z})^2)\end{aligned}$$

2 段階法で求めた形が十分な説明力を持っているかどうかの問題を過剰識別制約の問題という。もし十分な説明力があれば、2 段階法の誤差 $\hat{\mathbf{e}}$ を操作変数 $\mathbf{Z}$ で回帰しても係数が

すべて 0 になるはずである。それを調べる問題を考えてみる。

回帰係数を $\mathbf{d}$ 、誤差を $\mathbf{u}_e$ として以下のような回帰式を考える。

$$\hat{\mathbf{e}} = \mathbf{Z}\mathbf{d} + \mathbf{u}_e \tag{4}$$

この回帰式で定数項を含むすべての係数が 0 になるかどうか調べる検定は RV と EV を回帰変動と誤差変動、df_{RV} と df_{EV} をそれらの自由度として、以下のように与えられる。

$$F = \frac{RV/df_{RV}}{EV/df_{EV}} \sim F_{df_{RV}, df_{EV}} \tag{5}$$

ここで、$\hat{\mathbf{e}} = \mathbf{Z}\hat{\mathbf{d}} + \hat{\mathbf{u}}_e$ とすると、誤差変動と回帰変動は以下で与えられる。

$$\begin{aligned} EV &= \hat{\mathbf{u}}_e'\hat{\mathbf{u}}_e = (\hat{\mathbf{e}} - \mathbf{Z}\hat{\mathbf{d}})'(\hat{\mathbf{e}} - \mathbf{Z}\hat{\mathbf{d}}) = (\hat{\mathbf{e}} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{e}})'(\hat{\mathbf{e}} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{e}}) \\ &= \hat{\mathbf{e}}'(\mathbf{I} - \mathbf{P}_\mathbf{Z})(\mathbf{I} - \mathbf{P}_\mathbf{Z})\hat{\mathbf{e}} = \hat{\mathbf{e}}'(\mathbf{I} - \mathbf{P}_\mathbf{Z})\hat{\mathbf{e}} \\ RV &= \hat{\mathbf{e}}'\hat{\mathbf{e}} - EV = \hat{\mathbf{e}}'\mathbf{P}_\mathbf{Z}\hat{\mathbf{e}} \end{aligned}$$

またこの RV の表現は最小 2 乗法で求めた結合仮説検定の統計量を使っても求められる。

$$\begin{aligned} (\mathbf{R}\hat{\mathbf{b}} - \mathbf{r})'[\mathbf{R}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\mathbf{b}} - \mathbf{r}) &= \hat{\mathbf{d}}'\mathbf{Z}'\mathbf{Z}\hat{\mathbf{d}} \\ &= ((\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{e}})'(\mathbf{Z}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{e}} \\ &= \hat{\mathbf{e}}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{e}} = \hat{\mathbf{e}}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{e}} = \hat{\mathbf{e}}'\mathbf{P}_\mathbf{Z}\hat{\mathbf{e}} \end{aligned}$$

但し、$\mathbf{R} = \mathbf{I}_{m+r+1}$， $\mathbf{r} = \mathbf{0}$ である。

最後に、(5)式を示しておく。

i) まず $\mathbf{Z}'\mathbf{Z}$ の正則性と対称性を用いると、ある正方行列 $\mathbf{F}$ で $\mathbf{Z}'\mathbf{Z} = \mathbf{F}'\mathbf{F}$ のように書けることから、$\mathbf{P}_\mathbf{Z}$ はある行列 $\mathbf{B}(N \times (m + r + 1))$ を用いて以下のように書ける。

$$\mathbf{P}_\mathbf{Z} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' = (\mathbf{Z}\mathbf{F}^{-1})(\mathbf{Z}\mathbf{F}^{-1})' = \mathbf{B}\mathbf{B}',$$

以下の関係を用いると

$$\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}} = \mathbf{X}(\boldsymbol{\beta} - \hat{\mathbf{b}}) + \boldsymbol{\varepsilon} = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z}]\boldsymbol{\varepsilon}$$

回帰変動 RV は以下のように書ける。

$$\begin{aligned} RV &= \hat{\mathbf{e}}'\mathbf{P}_\mathbf{Z}\hat{\mathbf{e}} = \boldsymbol{\varepsilon}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z}]'\mathbf{P}_\mathbf{Z}[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z}]\boldsymbol{\varepsilon} \\ &= \boldsymbol{\varepsilon}'[\mathbf{P}_\mathbf{Z} - \mathbf{P}_\mathbf{Z}\mathbf{X}(\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z}]\boldsymbol{\varepsilon} \\ &= \boldsymbol{\varepsilon}'[\mathbf{B}\mathbf{B}' - \mathbf{B}\mathbf{B}'\mathbf{X}(\mathbf{X}'\mathbf{B}\mathbf{B}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{B}\mathbf{B}']\boldsymbol{\varepsilon} \\ &= \boldsymbol{\varepsilon}'\mathbf{B}[\mathbf{I}_{m+r+1} - \mathbf{B}'\mathbf{X}(\mathbf{X}'\mathbf{B}\mathbf{B}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{B}]\mathbf{B}'\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}'\mathbf{B}[\mathbf{I}_{m+r+1} - \mathbf{P}_{\mathbf{B}'\mathbf{X}}]\mathbf{B}'\boldsymbol{\varepsilon} \end{aligned}$$

ここで、

$$\begin{aligned} \mathbf{B}'\boldsymbol{\varepsilon} &\sim N(\mathbf{0}, \sigma^2\mathbf{B}'\mathbf{B}) = N(\mathbf{0}, \sigma^2\mathbf{F}'^{-1}\mathbf{Z}'\mathbf{Z}\mathbf{F}^{-1}) \\ &= N(\mathbf{0}, \sigma^2\mathbf{F}'^{-1}\mathbf{F}'\mathbf{F}\mathbf{F}^{-1}) = N(\mathbf{0}, \sigma^2\mathbf{I}_{m+r+1}) \end{aligned}$$

であることから、以下となる。

$$\mathbf{z} \equiv \mathbf{B}'\boldsymbol{\varepsilon}/\sigma \sim N(\mathbf{0}, \mathbf{I})$$

また $(\mathbf{B}'\mathbf{X})((m+r+1)\times(k+r+1))$ で $k \le m$ であることから、

$$rank(\mathbf{P}_{\mathbf{B}'\mathbf{X}}) = k+r+1, \quad rank(\mathbf{I}_{m+r-1} - \mathbf{P}_{\mathbf{B}'\mathbf{X}}) = m-k$$

よって、【公式 4】より、

$$RV/\sigma^2 = \hat{\mathbf{e}}'\mathbf{P}_{\mathbf{Z}}\hat{\mathbf{e}}/\sigma^2 = \mathbf{z}'[\mathbf{I}_{m+r+1} - \mathbf{P}_{\mathbf{B}'\mathbf{X}}]\mathbf{z} \sim \chi^2_{m-k}$$

ii) 次に、$N \to \infty$ を仮定すると、

$$rank(\mathbf{I} - \mathbf{P}_{\mathbf{Z}}) = N-m-r-1$$

と【公式 4】より、

$$EV/\sigma^2 = (\hat{\mathbf{e}}/\sigma)'(\mathbf{I} - \mathbf{P}_{\mathbf{Z}})(\hat{\mathbf{e}}/\sigma) \xrightarrow[N\to\infty]{} (\boldsymbol{\varepsilon}/\sigma)'(\mathbf{I} - \mathbf{P}_{\mathbf{Z}})(\boldsymbol{\varepsilon}/\sigma) \sim \chi^2_{N-m-r-1}$$

iii) また、$\mathbf{P}_{\mathbf{Z}}\hat{\mathbf{e}}$ と $(\mathbf{I} - \mathbf{P}_{\mathbf{Z}})\hat{\mathbf{e}}$ は

$$\begin{aligned}
\mathbf{P}_{\mathbf{Z}}\boldsymbol{\Sigma}_{\hat{\mathbf{e}}}(\mathbf{I} - \mathbf{P}_{\mathbf{Z}}) &= \mathbf{P}_{\mathbf{Z}}[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}]\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}[\mathbf{I} - \mathbf{P}_{\mathbf{Z}}\mathbf{X}(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'](\mathbf{I} - \mathbf{P}_{\mathbf{Z}}) \\
&= \sigma^2\mathbf{P}_{\mathbf{Z}}[\mathbf{I} - \mathbf{P}_{\mathbf{Z}}\mathbf{X}(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}' - \mathbf{X}(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{Z}} \\
&\qquad + \mathbf{X}(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'](\mathbf{I} - \mathbf{P}_{\mathbf{Z}}) \\
&= \sigma^2\mathbf{P}_{\mathbf{Z}}(\mathbf{I} - \mathbf{P}_{\mathbf{Z}}) = \mathbf{0}
\end{aligned}$$

であるから、【公式 2】より、独立である。

これらのことから、以下の関係を得る。

$$F = \frac{RV/(m-k)}{EV/(N-m-r-1)} = \frac{(\hat{\mathbf{e}}'\hat{\mathbf{e}} - \hat{\mathbf{u}}_e'\hat{\mathbf{u}}_e)/(m-k)}{\hat{\mathbf{u}}_e'\hat{\mathbf{u}}_e/(N-m-r-1)} \underset{N\to\infty}{\sim} F_{m-k,\infty}$$

3.2　プログラムの利用法

メニュー［分析－多変量解析他－経済・経営手法－操作変数回帰分析］を選択すると、図 2 に示す操作変数回帰分析の実行メニューが表示される。

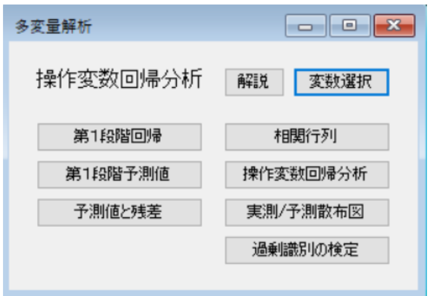

図 2 操作変数回帰分析実行メニュー

この分析用のデータは図 3 の形式で与えられる。

データ編集 操作変数回帰分析1.txt

	目的変数	@内生変数1	@内生変数2	$操作変数1	$操作変数2	外生変数
1	333	162	194	51	107	100
2	320	143	184	45	92	108
3	340	160	181	53	99	110
4	323	143	182	47	93	106
5	300	130	181	41	83	116
6	311	159	174	50	103	86
7	308	130	172	42	86	109
8	296	132	169	44	86	103
9	316	135	182	46	89	112

1/1 (1,1) 分析 備考

図 3 操作変数回帰分析のデータ形式

操作変数回帰分析の変数選択では、最初に目的変数を選択するが、その後の順番は特に決まっていない。しかし、内生変数の変数名の先頭には「@」記号、操作変数の変数名の先頭には「$」記号を付けて区別する。先頭にこれらの記号が付いていない変数は外生変数と解釈される。ここではすべての変数を選択すると、内生変数２つ、外生変数１つ、操作変数２つのモデルである。

まず、「第１段階回帰」ボタンをクリックして、操作変数の妥当性を調べる。操作変数の妥当性は第１段階回帰での F 値が 2 以上のとき妥当であると判断される。結果を図 4 に示す。

第１段階回帰分析結果

@内生変数1	偏回帰係数	標準化係数	標準誤差	z統計量	確率値	95%下限	95%上限	相関係数
$操作変数1	1.5112	0.4555	0.2140	7.0627	0.0000	1.0919	1.9306	0.9541
$操作変数2	0.8758	0.5747	0.1366	6.4120	0.0000	0.6081	1.1435	0.9663
外生変数	0.0325	0.0181	0.1065	0.3054	0.7601	-0.1762	0.2412	-0.6825
切片	-11.2521	0.0000	16.2512	-0.6924	0.4887	-43.1039	20.5998	
R	0.989	R^2	0.978	調整済R	0.987	調整済R^2	0.973	
F値	439.9434	自由度	3,∞	検定確率	0.0000			
@内生変数2	偏回帰係数	標準化係数	標準誤差	z統計量	確率値	95%下限	95%上限	相関係数
$操作変数1	0.2155	0.1089	0.5647	0.3817	0.7027	-0.8913	1.3224	0.7436
$操作変数2	0.9600	1.0559	0.2416	3.9733	0.0001	0.4865	1.4336	0.8199
外生変数	0.5024	0.4699	0.2315	2.1704	0.0300	0.0487	0.9560	-0.3481
切片	28.2781	0.0000	35.6531	0.7931	0.4277	-41.6006	98.1569	
R	0.885	R^2	0.783	調整済R	0.862	調整済R^2	0.742	
F値	28.1095	自由度	3,∞	検定確率	0.0000			

図 4 第１段階回帰分析結果

２つの内生変数に対して、それぞれ２つの操作変数と１つの外生変数による回帰式が示されている。ここでの標準誤差やそれに基づく指標の計算には、不均一分散を考慮した方法を用いている。この場合、操作変数の妥当性は満たされている。

「第１段階予測値」ボタンをクリックすると内生変数ごとの第１段階回帰分析の予測値が図 5 のように表示される。この値が第２段階の操作変数回帰分析に使われる。

第 1 段階回帰予測値

	@内生変数1	@内生変数2
1	162.780	192.228
2	140.836	180.553
3	159.121	190.002
4	144.669	180.939
5	127.169	175.070
6	157.310	181.139
7	131.080	174.649
8	133.908	172.066
9	139.850	179.898

図 5 第 1 段階回帰分析の予測値

操作変数回帰分析では、内生変数の実測値の代わりに第 1 段階回帰分析の予測値が用いられる。結果を図 6 に示す。

操作変数回帰分析結果

目的変数	偏回帰係数	標準化係数	標準誤差	z統計量	確率値	95%下限	95%上限	相関係数
@内生変数1	0.2538	0.2599	1.0072	0.2520	0.8011	-1.7202	2.2278	0.700
@内生変数2	1.4342	0.8763	1.3737	1.0440	0.2965	-1.2583	4.1267	0.735
外生変数	0.6470	0.3698	0.8311	0.7785	0.4363	-0.9819	2.2760	-0.113
切片	-52.5328	0.0000	87.7868	-0.5984	0.5496	-224.5917	119.5262	
R	0.763	R^2	0.582	調整済R	0.709	調整済R^2	0.503	
F値	14.476	自由度	3,∞	検定確率	0.000			

図 6 操作変数回帰分析の結果

ここでの標準誤差及びそれに基づく指標の導出には、操作変数回帰分析のパラメータと内生変数の実測値を用いた予測値からの残差を利用し、不均一分散を考慮した方法を用いている。

「予測値と残差」ボタンをクリックすると、操作変数回帰分析から推定されたパラメータの値と内生変数の実測値を用いて計算された予測値と残差の値が図 7 のように表示される。

予測値と残差

	実測値	予測値	残差
1	333	331.516	1.484
2	320	317.528	2.472
3	340	318.834	21.166
4	323	313.366	9.634
5	300	315.103	-15.103
6	311	293.012	17.988
7	308	297.666	10.334
8	296	289.989	6.011
9	316	315.218	0.782

図 7 操作変数回帰分析の予測値と残差

この予測値と実測値の関係は「実測/予測散布図」ボタンをクリックすることで図 8 のような散布図として表示される。ここで、グラフは重回帰分析などと同じく、縦軸が実測値、横軸が予測値である。

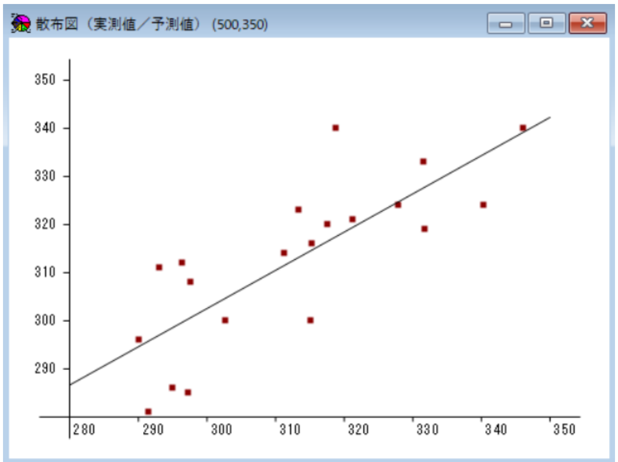

図 8 実測/予測散布図

操作変数の外生性の問題については、「過剰識別の検定」ボタンで調べることができる。但し、内生変数数より操作変数が多い場合のみ検定可能であるため、この例のように、内生変数２つ、操作変数２つの場合は検定を行うことができない。このソフトではこの条件を満たさない場合は、図 9 のような結果が表示される。

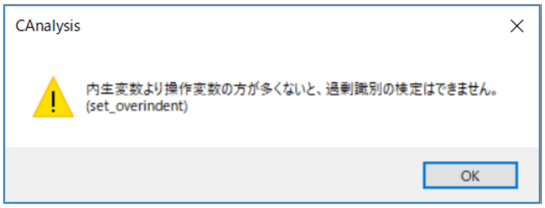

図 9 過剰識別の検定の注意メッセージ

今の場合、例えば、内生変数を 1 つ減らして内生変数 1 だけにすれば調べることが可能となる。結果は図 10 のように表示される。特に最下行の「過識別 F 値」のところが過剰識別の検定の部分である。「過識別確率」の値が有意水準の確率より大きければ（例えば 0.05 より大きければ）操作変数の外生性に問題はないと考える。

過剰識別の検定結果

IV回帰残差	偏回帰係数	標準化係数	標準誤差	z統計量	確率値	95%下限	95%上限	相関係数
$操作変数1	-1.1211	-0.6740	0.8429	-1.3301	0.1835	-2.7731	0.5309	-0.084
$操作変数2	0.5480	0.7174	0.5087	1.0773	0.2813	-0.4490	1.5451	0.067
外生変数	0.0660	0.0735	0.2958	0.2232	0.8234	-0.5137	0.6457	0.000
切片	-5.7737	0.0000	53.2953	-0.1083	0.9137	-110.2307	98.6832	
R	0.324	R^2	0.105	調整済R		調整済R^2		
F値	0.596	自由度	3,∞	検定確率	0.618			
過識別F値	0.888	過識別自由…	2,∞	過識別確率	0.412	≧αで良		

図 10 過剰識別の検定結果

4. おわりに

分析ソフトウェア C.Analysis は保健分野の統計学から始まっており、経済分野独特の

分析手法はあまり取り扱って来なかった。今回分野を広く見直す機会を得て、経済分野の考え方や分析手法を取り入れることにした。データ収集の困難さから、我々自身これらの分析を頻繁に利用することはないと思われるが、経済学部に通う学生が既存のデータを用いて気軽に分析を試し、特徴を直感的に理解してくれることを期待する。

この論文ではパネルデータ分析と操作変数回帰分析を取り上げた。過去に C.Analysis ではパネル重回帰分析と名付けた分析を導入したが、これは過去の変動のパターン認識と重回帰分析を組み合わせたもので、点予測に徹しており、パラメータの信頼区間などはほとんど検討していなかった。しかし、経済分析では回帰式の誤差項の不均一性や時系列的な相関を考慮し、予測よりも因果関係の把握を重視して、係数や予測の信頼区間を厳しく問う。ここで取り上げたパネルデータ分析もその特徴を強く持っている。さらに、操作変数回帰分析は操作変数をどのように収集するかが大きな問題であり、簡単に利用できる分析ではなさそうである。しかし、これらは経済学者の真の因果関係を見出そうとする真摯な態度を感じる手法である。

経済分野の組み込むべき手法で残っているのは、回帰分析を基にした時系列分析の手法である。これらは理論としてもプログラム開発としても難しい。特に、初心者にも使い易く簡単な操作で、というところは大きな課題である。

参考文献

［1］　J.H.Stock,M.W.Watson,宮尾龍蔵訳, 「入門計量経済学」, 共立出版, 2016.

補遺　行列と確率分布の公式

ここではこの論文で使う行列と分布の公式をまとめておく。$\mathbf{u}$ は確率変数である。

【公式１】　$Cov(\mathbf{Au},\mathbf{Bu}) = \mathbf{A}\boldsymbol{\Sigma}_\mathbf{u}\mathbf{B}'$

【公式２】　$\mathbf{A}\boldsymbol{\Sigma}_\mathbf{u}\mathbf{B}' = \mathbf{0}$ ならば、$\mathbf{Au}$ と $\mathbf{Bu}$ は独立した分布である。

【公式３】　$\mathbf{u}(m \times 1) \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{u})$ のとき、

$$\mathbf{d} + \mathbf{Au} \sim N(\mathbf{d}, \mathbf{A}\boldsymbol{\Sigma}_\mathbf{u}\mathbf{A}'), \quad \mathbf{u}'\boldsymbol{\Sigma}_\mathbf{u}^{-1}\mathbf{u} \sim \chi_m^2$$

【公式４】　$\mathbf{u} \sim N(0, \mathbf{I}_m)$ で $\mathbf{C}(m \times m)$ がべき等行列（$\mathbf{CC} = \mathbf{C}$）のとき、

$$\mathbf{u}'\mathbf{Cu} \sim \chi_r^2 \quad \text{但し、} rank(\mathbf{C}) = r$$

Multi-purpose Program for Social System Analysis 36
- Panel Data Analysis and Instrumental Variables Regression Analysis -

Masayasu FUKUI[*1] and Shoko TONAI [*1]

**1 Department of Business Administration, Faculty of Business Administration,*
Fukuyama Heisei University

Abstract: We have been constructing a unified program on the social system analysis for purpose of education. Recently, we are developing our program aiming to incorporate econometric analysis approach. In this paper we introduce programs of panel data analysis and instrumental variables regression analysis to eliminate missing variable biases. Both programs are designed to be easy to use for beginners.
Key Words: panel data analysis, instrumental variables regression analysis, College Analysis, statistics