ここで

$$S_{wx} = \begin{bmatrix} \tilde{S}_{w1x} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \tilde{S}_{w2x} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \tilde{S}_{wGx} \end{bmatrix},$$

$$\tilde{S}_{w_g\mathbf{x}} = \left[\frac{1}{N}\sum_{i=1}^{N} \mathbf{w}_{gi1}\mathbf{x}'_{i1}, \frac{1}{N}\sum_{i=1}^{N} \mathbf{w}_{gi2}\mathbf{x}'_{i2}, \cdots \frac{1}{N}\sum_{i=1}^{N} \mathbf{w}_{giT}\mathbf{x}'_{it} \right],$$

$$S_{xy} = \begin{bmatrix} \mathbf{S}_{xy1} \\ \mathbf{S}_{xy2} \\ \vdots \\ \mathbf{S}_{xy\alpha} \end{bmatrix},$$

$$\underset{TK\times 1}{S_{xy_g}} = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_{i1}y_{gi1} \\ \vdots \\ \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_{iT}y_{git} \end{bmatrix}$$

G3SLS は漸近的に最小距離推定に一致する。G3SLS も 3SLS も一致推定量ではあるが、誤差項の分散共分散に誤差構成要素が入っていると有効推定ではなくなる。

3.3 単一方程式推定

前節では同時方程式パネルデータ分析の考え方を示し、同時方程式体系全体の情報を有効に使うことで、はじめて有効一致推定を得ることができると論じた。しかし、現実的に複数の連立方程式に複雑な誤差構成要素を取り込んで計算することは極めてやっかいなことであり、現実のパネルデータにおける同時方程式の推計は次のように簡便化した 2 段階で行っている。

1. 観察不可能な（latent variable）効果を一階の階差を取って消去する。
2. 内生変数に対して操作変数を見つけて 2SLS 推計する。この場合、操作変数は時間とともに変化する変数を用いる。

次のようなモデルを考えよう。

$$y_{it1} = \alpha_1 y_{it2} + z_{it1}\beta_1 + \alpha_{i1} + v_{it1} \tag{3.5}$$

$$y_{it2} = \alpha_2 y_{it1} + z_{it2}\beta_2 + \alpha_{i2} + v_{it2} \tag{3.6}$$

ここで z_{it1}, z_{it2} は外生変数、一般モデルでは固定効果 α_{i1} と α_{i2} は全ての説明変数と相関している。誤差項 v_{it1} と v_{it2} は z とは無相関である。y_{it2} は v_{it1} と、y_{it1} は v_{it2} と相関している。

(5) 式を推計する場合、$\alpha_{i1} + v_{it1}$ は全ての説明変数と相関しているので OLS 推計は不適切である。そこで α_{i1} を階差を取って消去し、プーリング 2SLS で推計する。

$$\triangle y_{it1} = \alpha_1 \triangle y_{it2} + \triangle z_{it1}\beta_1 + \triangle v_{it1} \tag{3.7}$$

この場合、誤差項 $\triangle v_{it1}$ は $\triangle z_{it1}$ とは無相関となる。

しかし $\triangle y_{it2}$ と $\triangle v_{it1}$ は相関している可能性があり、$\triangle y_{it2}$ に対して操作変数をあてがう必要がある。一般には z_{it2} に含まれていて z_{it1} に含まれていない変数であり、かつ時間とともに変化する変数を用いる。

このようにして 2 段階最小二乗法あるいは操作変数法によってパラメータ β_1 をなるべくバイアスを少なくするように推定するというのが定石である。ここでの特徴は操作変数を外から探してくるのではなく、すでに同時方程式体系の中に含まれている外生変数を操作変数として用いるということである。問題は誤差項が一つの確率変数で表わされているのではなく、誤差構成要素が複数あり、それぞれの誤差分布に配慮しなければならないということである。

この問題をより厳密にするために、次のような同時方程式モデルを考えよう[*4]。

$$y_1 = Z_1\delta_1 + u_1 \tag{3.8}$$

ここで $Z_1 = [Y_1, X_1], \quad \delta_1' = (\gamma_1', \beta_1')$

Y_1 は要素 g_1 の内生変数、X_1 は要素 k_1 の外生変数、$X = [X_1, X_2]$ は同時方程式体系共通の外生変数である。この方程式は X_2 の式から除外されている外生変数の数 k_2 が $g_1 - 1$ と同数かそれより大きいときに識別できる[*5]。

誤差構成要素を次のように仮定する。

$$u_1 = Z_\mu\mu_1 + v_1 \tag{3.9}$$

ここで $Z_\mu = (I_N \otimes \iota_T), \quad \mu_1' = (\mu_{11}, \ldots, \mu_{N1})$ と $v_1' = (v_{111}, \ldots, v_{NT1})$ は平均ゼロの確率変数である。

$$E\begin{pmatrix}\mu_1 \\ v_1\end{pmatrix}(\mu_1', v_1') = \begin{bmatrix} \sigma^2_{\mu_{11}} I_N & 0 \\ 0 & \sigma^2_{v_{11}} I_{NT} \end{bmatrix} \tag{3.10}$$

(8) は次のように変換できる。$Q = I_{NT} - P$, $P = I_N \otimes \bar{J}_T$,

$$Qy_1 = QZ_1\delta_1 + Qu_1 \tag{3.11}$$

$\tilde{y}_1 = Qy_1$, $\tilde{Z}_1 = QZ_1$ として (11) 式を 2SLS 推計する。その際 $\tilde{X} = QX$ を操作変数として用いる。

[*4] 以下の議論は Baltagi(2001、pp.111-15) に依拠している。

[*5] 識別のための必要条件は、モデル全体に含まれる外生（先決）変数 K から当該方程式に含まれる外生（先決）変数の数 k の差が、当該方程式に含まれる内生変数の数 g との間に次のような関係を持つことである。$K - k \geqq g - 1$

Within2SLS 推計は

$$\tilde{\delta}_{1with2SLS} = (\tilde{Z}_1' P_{\tilde{X}} \tilde{Z}_1)^{-1} \tilde{Z}_1' P_{\tilde{X}} \tilde{y}_1 \tag{3.12}$$
$$var(\tilde{\delta}_{1with2SLS}) = \sigma_{v_{11}}^2 (\tilde{Z}_1' P_{\tilde{X}} \tilde{Z}_1)^{-1}$$

Within2SLS は次の式を GLS 推計することによっても導出可能である。

$$\tilde{X}' \tilde{y}_1 = \tilde{X} \tilde{Z}_1 \delta_1 + \tilde{X}' \tilde{u}_1 \tag{3.13}$$

$\bar{y}_1 = Py_1, \bar{Z}_1 = PZ_1$ とおき $\bar{X} = PX$ を操作変数として (5) 式を 2SLS 推計すると Between2SLS 推計が得られる。

$$\hat{\delta}_{1btw2SLS} = (\bar{Z}_1' P_{\bar{x}} \bar{Z}_1)_1^{-1} \bar{Z}' P_{\bar{x}} \bar{y}_1 \tag{3.14}$$
$$var(\hat{\delta}_{1btw2SLS}) = \sigma_{1_{11}}^2 (\bar{Z}_1' P_{\bar{x}} \bar{Z}_1)^{-1}$$
$$\sigma_{1_{11}}^2 = T\sigma_{\mu_{11}}^2 + \sigma_{v_{11}}^2$$

Between2SLS は GLS 推計としても導出可能である。

$$\bar{X}' \bar{y}_1 = \bar{X} \bar{Z}_1 \delta_1 + \bar{X}' \bar{u}_1 \tag{3.15}$$

(13) と (15) を同時方程式として扱う。

$$\begin{pmatrix} \tilde{X}' \tilde{y}_1 \\ \bar{X}' \bar{y}_1 \end{pmatrix} = \begin{pmatrix} \tilde{X}' \tilde{Z}_1 \\ \bar{X}' \bar{Z}_1 \end{pmatrix} \delta_1 + \begin{pmatrix} \tilde{X}' \tilde{u}_1 \\ \bar{X}' \bar{u}_1 \end{pmatrix} \tag{3.16}$$

ここで

$$E \begin{pmatrix} \tilde{X}' \tilde{u}_1 \\ \bar{X}' \bar{u}_1 \end{pmatrix} = 0, \quad var \begin{pmatrix} \tilde{X}' \tilde{u}_1 \\ \bar{X}' \bar{u}_1 \end{pmatrix} = \begin{bmatrix} \sigma_{v_{11}}^2 \tilde{X}' \tilde{X} & 0 \\ 0 & \sigma_{1_{11}}^2 \bar{X}' \bar{X} \end{bmatrix}$$

(16) 式を GLS 推計すると the error component two-stage least squares (EC2SLS) を得る。

$$\hat{\delta}_{1,EC2SLS} = \left[\frac{\tilde{Z}_1' P_{\tilde{X}} \tilde{Z}_1}{\sigma_{v_{11}}^2} + \frac{\bar{Z}_1' P_{\bar{X}} \bar{Z}_1}{\sigma_{1_{11}}^2} \right]^{-1} \left[\frac{\tilde{Z}_1' P_{\tilde{X}} \tilde{y}_1}{\sigma_{v_{11}}^2} + \frac{\bar{Z}_1' P_{\bar{X}} \bar{y}_1}{\sigma_{1_{11}}^2} \right] \tag{3.17}$$

ここで

$$\hat{\delta}_{1EC2SLS} = w_1 \hat{\delta}_{1with2SLS} + w_2 \hat{\delta}_{1btw2SLS}$$
$$w_1 = \left[\frac{\tilde{Z}_1' P_{\tilde{X}} \tilde{Z}_1}{\sigma_{v_{11}}^2} + \frac{\bar{Z}_1' P_{\bar{X}} \bar{Z}_1}{\sigma_{1_{11}}^2} \right]^{-1} \left[\frac{\tilde{Z}_1' P_{\tilde{X}} \tilde{Z}_1}{\sigma_{v_{11}}^2} \right]$$
$$w_2 = \left[\frac{\tilde{Z}_1' P_{\tilde{X}} \tilde{Z}_1}{\sigma_{v_{11}}^2} + \frac{\bar{Z}_1' P_{\bar{X}} \bar{Z}_1}{\sigma_{1_{11}}^2} \right]^{-1} \left[\frac{\bar{Z}_1' P_{\bar{X}} \bar{Z}_1}{\sigma_{1_{11}}^2} \right]$$

$$\hat{\sigma}_{v_{11}}^2 = (y_1 - Z_1 \tilde{\delta}_{1with2SLS})' Q (y_1 - Z_1 \tilde{\delta}_{1with2SLS}) / N(T-1) \tag{3.18}$$
$$\hat{\sigma}_{1_{11}}^2 = (y_1 - Z_1 \tilde{\delta}_{1btw2SLS})' P (y_1 - Z_1 \hat{\delta}_{1btw2SLS}) / N \tag{3.19}$$
$$\hat{\sigma}_{\mu_{11}}^2 = (\hat{\sigma}_{1_{11}}^2 - \hat{\sigma}_{v_{11}}^2) / T > 0$$

この結果は、第1章でも繰り返し論じたように、EC2SLS推定はウイズイン推定とビトウィーン推定の加重平均になっている。

(8) 式に $\Omega_{11}^{-1/2}$ をかけて

$$y_1^* = Z_1^* \delta_1 + u_1^* \tag{3.20}$$

ここで $y_1^* = \Omega_{11}^{-1/2} y_1, \quad Z_1^* = \Omega_{11}^{-1/2} Z_1, \quad u_1^* = \Omega_{11}^{-1/2} u_1$

$$\Omega_{11}^{-1/2} = (P/\sigma_{1_{11}}) + (Q/\sigma_{v_{11}}) \tag{3.21}$$
$$y_{1_{it}}^* = (y_{1_{it}} - \theta_1 \bar{y}_{1i.})/\sigma_{v_{11}}, \quad \theta_1 = 1 - (\sigma_{v_{11}}/\sigma_{1_{11}})$$
$$\bar{y}_{1i.} = \sum_{t=1}^{T} y_{1it}/T$$

操作変数 A を用いて (20) 式を 2SLS 推計すると

$$\hat{\delta}_{1,2SLS} = (Z_1^{*'} P_A Z_1^*)^{-1} Z_1^{*'} P_A y_1^* \tag{3.22}$$

ここで $P_A = A(A'A)^{-1}A'$
最適操作変数を次のように表す。

$$X^* = \Omega_{11}^{-1/2} X = \frac{QX}{\sigma_{v_{11}}} + \frac{PX}{\sigma_{1_{11}}} = \frac{\tilde{X}}{\sigma_{v_{11}}} + \frac{\bar{X}}{\sigma_{1_{11}}}$$

$A = X^*$ とすると G2SLS を得る。

$$\hat{\delta}_{1,G2SLS} = (Z_1^{*'} p_{X^*}^* Z_1)^{-1} Z_1^* P_{X^*} y_1^* \tag{3.23}$$

(20) 式に操作変数 $A = [QX, PX] = \left[\tilde{X}, \bar{X}\right]$ を用いて 2SLS 推計を行う。ここで QX は PX に直交しており、$P_A = P_{\tilde{X}} + P_{\bar{X}}$ である。

$$\begin{aligned} P_A Z_1^* &= (P_{\tilde{X}} + P_{\bar{X}}) \left[\Omega_{11}^{-1/2} Z_1\right] \\ &= (P_{\tilde{X}} + P_{\bar{X}}) \left[\frac{Q}{\sigma_{v_{11}}} + \frac{P}{\sigma_{\ell_{11}}}\right] Z_1 = \frac{P_{\tilde{X}} \tilde{Z}_1}{\sigma_{v_{11}}} + \frac{P_{\bar{X}} \tilde{Z}_1}{\sigma_{1_{11}}} \end{aligned} \tag{3.24}$$

ここで

$$Z_1^{*\prime} P_A Z_1^* = \left(\frac{\tilde{Z}_1' P_{\tilde{X}} \tilde{Z}_1}{\sigma_{v_{11}}^2} + \frac{\bar{Z}_1' P_{\bar{X}} \bar{Z}_1}{\sigma_{1_{11}}^2} \right)$$
$$Z_1^{*\prime} P_A y_1^* = \left(\frac{\tilde{Z}_1' P_{\tilde{X}} \tilde{y}_1}{\sigma_{v_{11}}^2} + \frac{\bar{Z}_1' P_{\bar{X}} \bar{y}_1}{\sigma_{1_{11}}^2} \right)$$

(17) 式の $\hat{\delta}_{1,EC2SLS}$ は $A = \left[\tilde{X}, \bar{X}\right]$ の時 (23) 式と同値である。

すなわち、$EC2SLS$ は既存の 2SLS 推計法を用いて推計することが可能なのである。

第1ステップ：(8) 式 Within2SLS と (15) 式 Between2SLS を 2SLS で推計し、(12) と (14) 式を得る。

第2ステップ：$\hat{\sigma}^2_{v_{11}}$ と $\hat{\sigma}^2_{1_{11}}$ を (18)(19) によって推計し、(22) 式で用いる y_1^*, Z_1^*, X^* を得る。(8) 式を $\Omega_{11}^{-1/2}$ で変換して (20) 式を得る。

第3ステップ：操作変数 $A = X^*$ か $A = [QX, PX]$ を用いて (20) 式を 2SLS 推計すると、それぞれ (22) と (17) 式を得る。

3.4 内生性検定

内生性検定としては一般には、Wu-Hausman 検定として知られているもの（最小二乗法推定と操作変数法推定のパラメータをハウスマン検定する）や内生変数に関するモデルを最小二乗推定し、その式から得られた誤差をもともとの式に代入し、そのパラメータが 0 かどうかを t 検定するという方法が提案されている。

次のようなモデルで説明変数 y_2 の内生性の疑いがある時を考えよう[*6]。

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1$$

ここで z_1 と z_2 は外生変数であり、他に操作変数として z_3 と z_4 を考えることができる。この時、上の式を最小二乗法と操作変数法で推計し、パラメータが有意に違うかどうかを Hausman (1978) に従ってカイ二乗検定することによって、説明変数 y_2 が内生であるかどうかを確かめることができる[*7]。これは Durbin-Wu-Hausman (DWH) 検定として知られているが、基本的には Hausman 検定を敷衍したものである。すなわち、最小二乗法推定による推定パラメータ $\widehat{\beta}_e$ と操作変数法による推定パラメータ $\widehat{\beta}_c$ を用いて、次のようなカイ二乗統計量を計算する。

$$(\widehat{\beta}_c - \widehat{\beta}_e)'(var[\widehat{\beta}_c] - var[\widehat{\beta}_e])^{-1}(\widehat{\beta}_c - \widehat{\beta}_e) \backsim \chi(k_1)$$

ここで k_1 は内生性検定の対象となった内生変数の数である。この検定は内生性検定というよりも、最小二乗法と操作変数法を用いた場合に推定パラメータが有意に違うかどうかを検定したものである。

Wooldridge(2002, p.119) は次ような内生性検定法を紹介してる。ここで y_1 に関するモデルを考えよう。

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1$$

*6 以下の議論は北村 (2009)『ミクロ計量経済学入門』（日本評論社）第5章第6節を引用している。

*7 一般的には Durbin (1954)、Wu(1973)、Hausman(1978) によって形成された検定であり、Durbin-Wu-Hausman test として知られている。Bowden and Turkington (1984, pp.50-52) や Davidson and MacKinnon (2004, pp.338-340) を参照。

説明変数 y_2 が内生変数であると仮定して次のような式を推計する。

$$y_2 = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 z_3 + \alpha_4 z_4 + v_2$$

操作変数の仮定により z_j は u_1 とは無相関であるので、v_2 が u_1 と無相関であれば、y_2 も u_1 とは無相関になる。ということは次式でパラメター $\delta_1 = 0$ が y_2 も u_1 とは無相関のための必要十分条件になる。

$$u_1 = \delta_1 v_2 + e_1$$

これを直接検定する方法はないので、y_2 式を最小二乗法で推計し、残差として $\hat{v}_2$ を計算し、これをもとの y_1 式に代入し最小二乗法で推計する。

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta_1 \hat{v}_2 + \varepsilon$$

t 検定で $\delta_1 = 0$ が棄却されれば、y_2 は内生変数であるということになる。

ここでは、従来の内生性の問題とは違って、固定効果が内生であり、説明変数と相関している場合を考える[*8]。はじめに全ての説明変数が固定効果に相関している場合を考え、次いで一部の説明変数のみが固定効果と相関している場合を考えよう。

$$y = \alpha_{\iota NT} + X\beta + Z_\mu \mu + v = Z\delta + Z_\mu \mu + v \tag{3.25}$$
$$\mu_i = \bar{X}_i' \pi + \varepsilon_i \tag{3.26}$$

ここで $\varepsilon_i \backsim iid(0, \sigma_\varepsilon^2)$, $\bar{X}_i'$ は $1 \times K$ ベクトル
(26) は次のように書き換えられる。

$$\mu_i = Z_\mu' X\pi / T + \varepsilon_i \tag{3.27}$$

ここで $\mu' = (\mu_1, \ldots, \mu_N)$, $Z_\mu = I_N \otimes \iota_T$, $\varepsilon_i' = (\varepsilon_1, \ldots, \varepsilon_N)$.
(27) を (25) に代入すると、

$$y = X\beta + PX_\pi + (Z_\mu \varepsilon + v) \tag{3.28}$$

ここで $P = I_N \otimes \bar{J}_T$, ε と v は無相関で、$(Z_\mu \mu + v)$ は平均ゼロで次のような分散共分散行列構造をもつ。

$$V = E(Z_\mu \varepsilon + v)(Z_\mu \varepsilon + v)' = \sigma_\varepsilon^2 (I_N \otimes J_T) + \sigma_v^2 I_{NT} \tag{3.29}$$

[*8] 以下は Baltagi (2001,pp.118-122) を引用している。これは賃金関数において、個人の固定効果である能力と学歴や職歴が相関している場合、生産関数において潜在変数である経営能力が労働や資本などの投入財と相関している場合など様々なケースで出てくる問題である。

(28) の GLS 推計は次のようになる。

$$\hat{\beta}_{GLS} = \tilde{\beta}_{with} = (X'QX)^{-1}X'Qy \tag{3.30}$$

$$\hat{\pi}_{GLS} = \hat{\beta}_{btw} - \tilde{\beta}_{with} = (X'PX)^{-1}X'Py - (X'QX)^{-1}X'Qy \tag{3.31}$$

$$var(\hat{\pi}_{GLS}) = var(\hat{\beta}_{btw}) + var(\tilde{\beta}_{with}) \tag{3.32}$$

$$= (T\sigma_\varepsilon^2 + \sigma_v^2)(X'PX)^{-1} + \sigma_v^2(X'QX)^{-1} \tag{3.33}$$

Mundlak(1978) が示したように、(25) 式の最良線形不偏推定量 (BLUE) は固定効果 (within) 推定である。ランダム効果推定は (26) 式を無視しておりバイアスが残る。(28) 式では全ての説明変数は固定効果に相関しているが、ランダム効果モデルでは説明変数と固定効果は無相関であることが想定されている。

Hausman and Taylor(1981) では、一部の説明変数のみが μ_i と相関しているというモデルを考えている。

$$y_{it} = X_{it}\beta + Z_i\gamma + \mu_i + v_{it} \tag{3.34}$$

Hausman and Taylor (1981) は $X = [X_1; X_2]$ と $Z = [Z_1; Z_2]$ を 2 分割した。すなわち、X_1 は $n \times k_1$, X_2 は $n \times k_2$, Z_1 は $n \times g_1$ Z_2 は $n \times g_2$, $n = NT$ に分割し、X_1 と Z_1 は外生変数、X_2 と Z_2 は内生変数で μ_i と相関し、v_{it} とは無相関であるとする。

ウィズイン誤差項を次のように求める。

$$\hat{d}_i = \bar{y}_i - \bar{X}_i\hat{\beta}_w \tag{3.35}$$

(35) の時間平均をとり、$\hat{d}_i$ を z_i に関して、操作変数 $A = [X_1, z_1]$ を用いた 2 SLS 推計を行う。

$$\hat{\gamma}_{2SLS} = (Z'P_AZ)^{-1}Z'P_A\hat{d} \tag{3.36}$$

ここで $P_A = A(A'A)^{-1}A'$。$Z'P_AZ$ は非特異 (non singular) 行列であり、次数条件 $k_1 \geqslant g_2 - 1$ を満たしている。

分散は次のように求められる。

$$\hat{\sigma}_v^2 = \tilde{y}'\bar{P}_{\tilde{X}}\tilde{y}/N(T-1) \tag{3.37}$$

ここで $\tilde{y} = Qy$, $\tilde{X} = QX$, $\bar{P}_A = 1 - P_A$

$$\sigma_1^2 = \frac{(y_{it} - X_{it}\tilde{\beta}_w - Z_i\hat{\gamma}_{2SLS})'P(y_{it} - X_{it}\tilde{\beta}_w - Z_i\hat{\gamma}_{2SLS})}{N} \tag{3.38}$$

(35) を次のように変換する。

$$\Omega^{-1/2}y_{it} = \Omega^{-1/2}X_{it}\beta + \Omega^{-1/2}Z_1\gamma + \Omega^{-1/2}u_{it} \tag{3.39}$$

Hausman and Taylor 推定は、(39) 式を $A_{HT} = \left[\tilde{X}, \tilde{X}_1, Z_1\right]$ を操作変数とした 2SLS 推計であると理解できる。

1. $k_1 < g_2 - 1$ であれば過小識別、$\hat{\beta}_{HT} = \tilde{\beta}_{with}$ であり $\hat{\gamma}_{HT}$ は存在しない。
2. $k_1 = g_2 - 1$ であれば適正識別、$\hat{\beta}_{HT} = \hat{\beta}_{with}$, $\hat{\gamma}_{HT} = \hat{\gamma}_{2SLS}$ である。
3. $k_1 > g_2 - 1$ であれば過剰識別、(39) 式より得られた $\hat{\beta}_{HT}$ は $\hat{\beta}_{with}$ より有効である。

過剰識別テストは次の統計量によってテストできる。

$$\hat{m} = \hat{q}' \left[var(\tilde{\beta}_{with}) - var(\hat{\beta}_{HT}) \right]^{-1} \hat{q} \tag{3.40}$$

ここで $\hat{q} = \hat{\beta}_{HT} - \tilde{\beta}_{with}$, $\hat{\sigma}_v^2 \hat{m} \overset{H_0}{\to} X_\ell^2$, $\ell = \min[k_1 - q_2, NT - k]$ である。

3.5 不均一分散検定

不均一分散の問題は操作変数法にも残っており、検定を行い、不均一分散の問題を解決することが望ましい[*9]。また、パラメータの分散を不均一分散頑強標準誤差によって修正し、頑強 t 統計量を推定する必要がある。基本的な考え方は第 4 章で見た、Breusch and Pagan (1979) の検定と White (1980) の不均一誤差頑強推定を踏襲するものである。

しかし、Pagan and Hall (1983) が指摘したように、複数の内生変数をもつ操作変数法を考えるような一般的設定では、Breusch and Pagan (1979) の検定は、関心のある内生変数を含んだ式のみの不均一分散を検定しており、潜在的な他の連立方程式における不均一分散問題は無視している。Pagan and Hall (1983) および White (1982) は他の連立方程式に不均一分散問題が存在しているという設定でカイ二乗検定を提案している[*10]。

3.6 弱相関の操作変数の問題

実証研究上、適切な操作変数を見つけることは極めて難しいことが知られている[*11]。とりわけ z と x の相関が弱い場合には問題がある。変数 z と誤差項 u が相関している場合の操作変数法による推計値の確率極限は次のように表せる。

$$p\lim \hat{\beta}_1 = \beta_1 + \frac{Corr(z,u)}{Corr(z,x)} \cdot \frac{\sigma_u}{\sigma_x}$$

σ_u と σ_x は、u と x に関する標準偏差である。問題は例え、$Corr(z,u)$ が小さくても、$Corr(z,x)$ も小さければ操作変数による推計値 $\hat{\beta}_1$ は大幅な不一致推定となるということである。現実的に

[*9] 以下の議論は北村 (2009)『ミクロ計量経済学入門』(日本評論社) 第 5 章第 4 節を引用している。
[*10] STATA では ivhettest というコマンドを使うことで、Pagan and Hall (1983)、White (1980)、Breusch and Pargan (1979)、Koenker (1981) などの一連の不均一分散検定を行うことができる。
[*11] 以下の議論は北村 (2009)『ミクロ計量経済学入門』(日本評論社) 第 5 章第 5 節を引用している。

考えて、操作変数を用いるよりも最小二乗法を用いた方が不一致性の程度が低くなることもあり得る。

これは**弱相関操作変数** (Weak Instrumental Variables) の問題として知られている。以下ではこの弱相関に関する主要な検定を紹介する。

Bound, Jaeger and Baker (1995) は内生変数を操作変数で回帰した第1段階の推定式の決定係数 (R^2) の計算において、いくつかの操作変数を落として推定することによって決定係数に変化があるかどうかを検定することを提案している。具体的には部分決定係数は $(RSS_{Z_2} - RSS_Z)/TSS$ と定義され、RSS_{Z_2} は操作変数 Z_2 だけを使って計算した誤差平方和であり、RSS_Z はすべての操作変数を用いた場合の誤差平方和を表す。この方法は弱相関の問題を検知するが、その結果として内生変数に対して操作変数が過小になる場合を排除できない。

Shea(1997) は操作変数間の相関を考慮した部分決定係数を提案している。すなわち、内生変数 i に関する部分決定係数は $R_p^2 = (v_{i,i,OLS})/(v_{i,i.IV})\{(1 - R_{IV}^2)/(1 - R_{OLS}^2)\}$ と表される。ここで $v_{i,i}$ は内生変数にかかる推定係数の漸近分散を表している[*12]。

Andreson(1984) とそれを敷衍した Hall, Rudebusch and Wilcox (1996) はより一般的なアプローチを提案しいてる。ここでは内生変数 $\mathbf{X}$ と操作変数 $\mathbf{Z}$ の行列間の**正準相関** (canonical correlation) $Corr_i$ $i = 1, 2, ..k$ を計算し、操作変数が有意であるということは、すべての相関が有意にゼロとは異なるはずであることを検定している。Anderson は最小の正準相関はゼロであるという帰無仮説を尤度比を用いて検定した。この統計量は自由度 $l - k + 1$ のカイ二乗分布に従う。帰無仮説が棄却できなければ操作変数の弱相関問題だけではなく、識別に問題がある可能性を示唆することになる。

Hall and Peixe (2000) は正準相関を用いて操作変数の重複 (redundancy) を検定する方法を提案した。これは Anderson の検定に似ているが、重複していると疑われる操作変数を含んだ正準相関と含まない正準相関に関する尤度比検定であり、自由度が内生変数と重複している操作変数の積に等しくなるようなカイ二乗分布に従うことを示した[*13]。

操作変数の弱相関問題は、多くの無相関あるいは低相関の操作変数を用いると最小二乗法推定以上にバイアスをもたらすことが Hahn and Hausman (2002b) によって示されている。また、Staiger and Stock (1997) は弱相関問題は第1段階の推定において操作変数が有意であっても起こりうることを示している[*14]。

[*12] Shea(1997) の統計量は小さい程、操作変数の内生変数に関する説明力が低いことを意味しいてる。STATA のコマンド ivreg2 の中の first か ffirst というオプションを用いれば計算できる。

[*13] この検定は STATA では ivreg2 の中の redundant というオプションを用いればいい。また ivreg2 の中にでは Anderson and Rubin (1949)、Cragg and Donald (1993)、Stock and Wright (2000) らの検定が出来る。

[*14] Nelson and Sartz (1990)、Stock and Wright (2000)、Stock, Wright and Yogo (2002)、Hahn and Hausman (2002a, 2003)、Andrews and Stock (2005)、Chao and Swanson (2005)、Stock and Yogo (2005)、Hausman, Stock and Yogo (2005) なども参照。

3.7 STATA コード

ここでは『21 世紀出生児縦断調査』を用いて操作変数法パネル推定を行ってみよう。これまで北村 (2007、2008, 2009) で行ってきた出生児の身体成長パターンの測定を行ってきた。これまでの研究では操作変数パネル推定ではなく、身長・体重の対数を出生からの経過日数とその二乗、および子育費用によって説明する固定効果モデルを採用してきた。

しかし、第 2 章でモデルに一次のラグ項を入れた推定を行うと推定係数に内生性バイアスがあるように見受けられた。そこで、本章では子育費用の対数値を被説明変数とするモデルを考えてみたい。基本的な考え方は、子育費用は母親がフルタイムの職に就いていてそれなりの所得を得ているかどうかに依存するが、母親の就業は父親の所得とベビーシッターの利用可能性に依存しているというものである。

$$\ln costofchildcare_{it} = \alpha + \gamma fulltimework\_m_{it} + \beta workdummy\_m_{it} + u_{it}$$
$$workdummy\_m_{it} = \delta \ln income\_f + \eta childcarewor\,ker + e_{it}$$

ここで $\ln costofchildcare$ = 子育費用の対数値、$fulltimework\_m$ = 母親が週 40 時間以上の労働に従事している場合に 1 をとるダミー変数、$workdummy\_m$ = 母親の就業状況を表すダミー変数。フルタイム、パートタイム、自営業、内職などいずれかの職についていれば 1 をとる。$\ln income\_f$ は父親の所得の対数値、$childcarewor\,ker$ は普段の保育者が保育ママやベビーシッターであれば 1 をとるダミー変数。

推計に使った STATA コードは次のようになる。

```
/*weak instruments tests*/
ivreg2 lncostofchildcare fulltimework_m (workdummy_m = lnincome_f childcareworker),
gmm2s orthog(lnincome_f)
ivhettest, all
/*Durbin-Wu-Hausman tests for endogeneity in Iv estimation*/
quietly ivreg2 lncostofchildcare fulltimework_m (workdummy_m = lnincome_f childcare-
worker), small
estimates store iv2
quietly regress lncostofchildcare fulltimework_m workdummy_m
hausman iv2, constant sigmamore
quietly ivreg2 lncostofchildcare fulltimework_m (workdummy_m = lnincome_f childcare-
worker), orthog(lnincome_f) small
ivendog
/*Panel Instrumental Variable Estimation*/
xtivreg lncostofchildcare fulltimework_m (workdummy_m = lnincome_f childcare-
```

```
worker), fe
est store fixed4
xtivreg lncostofchildcare fulltimework_m (workdummy_m = lnincome_f childcare-
worker), re
hausman fixed4
```

まず、図表 3.1 では 3.6 節で論じた操作変数の有意性検定を行った。ここではデータをプールして検定している。各検定の統計的意義や解釈については北村 (2009) を参照されたい。ここでは、Anderson Canonical Correlation 尤度比検定では帰無仮説が棄却できることで、操作変数が有意であることを示唆している。Cragg and Donald 検定でも弱相関問題が棄却されていること、そして、不均一分散検定に関しては、Pagan and Hall 検定、White/Koenker 検定、Breusch and Pagan 検定などがあり、不均一分散の存在が棄却できないことを示唆していることを報告しておきたい。

Dependent Variable: lncostofchildcare	2step GMM Estimated Coefficient	z-stat
workdummy_m	0.767	76.77
fulltimework_m	−0.008	−0.74
_cons	0.660	171.83
Number of observation	142886	
Centerd R2	−0.012	
Uncentered R2	0.549	
Root MSE	0.840	
Identification Tests		
Underidentification test (Anderson canonical Correlation LM statistic)	Chi2(2) = 3.6e+04 P-value = 0.0000	
Weak identification test (Cragg-Donald wald F statistic)	2.4e + 04	
Sargan statistic	Chi2(1) = 2188.256 P-value = 0.0000	
C Statistic	Chi2(1) = 2188.256 P-value = 0.0000	
IV heteroskedasticity tests		
Pagan-Hall general test statistic	Chi2(3) = 377.345 P-value = 0.0000	
Pagan-Hall test	Chi2(3) = 801.917 P-value = 0.0000	
White/Koenker nR2 test statistic	Chi2(3) = 365.547 P-value = 0.0000	
Breusch-Pagan/Godfrey/Cook-Weisberg	Chi2(3) = 805.520 P-value = 0.0000	
Instrumented:	Workdummy_m	
Included instruments:	lnincome_m fulltimework_m	
Excluded instruments:	lnincome_f childcareworker	

図表 3.1 操作変数の有意性検定

Dependent variable: Incostofchildcare	Coefficients (b)iv	(B)	(b-B)	sqrt(diag(V_b-V_B))
workdummy_m	0.767	0.264	0.503	0.008
fulltimework_m	-0.008	0.222	-0.230	0.008
_cons	0.660	0.858	-0.198	0.003

Test Ho: Difference in coefficients not systematic: Chi2(3) = 4823.29, Prob>chi2 = 0.0000

Tests of endogeneity of workdummy_m
Wu-Hausman F test: 3.20e+03 F(1, 142882) P-value = 0.0000
Durbin-Wu-Hausman chi2 test: 3.13e+03 Chiw(1) P-value = 0.0000

図表 3.2 内生性検定

Dependent Variable: Incostofchildcare	Fixed effect Estimated Coefficient	z-statistics	Random Effect Estimated Coefficient	z-statistics
workdummy_m	1.477	66.17	0.825	77.19
fulltimeworker_m	0.080	6.44	-0.010	-0.90
_cons	0.401	50.04	0.638	152.31
Diagnostic Tests				
Number of observation	142886		142886	
Number of groups	45214		45214	
R-sq:				
within	-		0.023	
between	0.053		0.052	
overall	0.043		0.042	
F-test that all u_i=0	F(45213, 97670) = 1.34, Prob > F = 0.0000			
Hausman test	Chi2(2) = 1533.34, Prob>chi2 = 0.0000			
Instrumented:	workdummy_m			
Instruments:	fulltimework_m lnincome_f childcareworker			

図表 3.3 パネル操作変数方推定

図表 3.2 は 3.4 節で論じた内生性検定を行った。Hausman 検定、Wu-Hausman 検定、Durbin-Wu-Hasuman 検定全てで、母親の就業ダミーの外生性が棄却されている。

図表 3.3 では、母親の就業を内生変数として扱った、子育費用の操作変数パネル推定を行った。結果として固定効果推定が選択され、母親の就業ダミーは有意に正の効果を持つことが確認された。

本章では同時方程式パネル推定を行ったが、『21 世紀出生児縦断調査』で継続的に同じ質問を繰り返して、データを蓄積するというパネルデータ特有のデータが極めて限定されている。さらに、操作変数として使えるような変数はさらに限定的であることから、この推定方法が有効に用いられるデータ環境には現状ではないと言わざるを得ない。

参考文献

[1] 北村行伸 (2005) 『パネルデータ分析』、岩波書店

[2] 北村行伸 (2009) 『ミクロ計量経済学入門』、日本評論社

[3] Anderson, T.W. (1984) *Introduction to Multivariate Statistical Analysis*, Wiley.

[4] Anderson, T.W. and Rubin, H.(1949) "Estimators of the Paramters of a Single Equation in a Complete Set of Stochastic Equations", *Annals of Mathematical Statistics*, 21, pp.570-82.

[5] Andrew, Donald W.K. and Stock, James H.(2005) "Inference with Weal Instruments", NBER Technical Working Paper 313.

[6] Angrist, J.D.and Krueger, A.B.(1991) "Does Compulsory School Attendance Affect Schooling and Earnings?", *Quarterly Journal of Economics*, 106, pp.979-1014.

[7] Angrist, Joshua D. and Krueger, Alan B.(2001) "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments", *Journal of Economic Perspectives*, 15(4), pp.69-85.

[8] Baltagi, B.H. (1981b) "Simultaneous Equations with Error Components," *Journal of Econometrics*, 17, pp.189-200.

[9] Baltagi, B.H. (2001) *Econometric Analysis of Panel Data*, 2nd ed, John Wiley & Sons.

[10] Basmann, R.L. (1960) "On Finite Sample Distributions of Generalized Classical Linear Identifiability Test Statistics", *Journal of the American Statistical Association*, 55(292), pp.650-59.

[11] Baum, Christopher. (2006) *An Introduction to Modern Econometrics Using Stata*, Stata Press.

[12] Blackburn, McKinley and Neumark, David.(1992) "Unobserved Ability, Efficiency Wages, and Interindustry Wage Differentials", *Quaterly Journal of Economics*, 107(4), pp.1421-1436.

[13] Bound, John., Jaeger, David.A. and Baker, Regina. M.(1995) "Problems with Instrumental Variables Estimation when the Correlation between the Instruments and the Endogenous Explanatory Variable is Weak", *Journal of the American Statistical Association*, 90(430), pp.443-50.

[14] Bowden, R.J. and Turkington, D.A.(1984) *Instrumental Variables*, Cambridge University

Press.

[15] Breusch, Trevor., Qian, Hailong., Schmidt, Peter., and Wyhowski, Donald.(1999) "Redundancy of Moment Conditions", *Journal of Econometrics*, 91, pp.89-111.

[16] Chamberlain, G. (1977) "Education, Income, and Ability Revisited," in *Latent Variables in Socio-Economic Models*, eds. by D.J. Aigner and A.S. Goldberger, pp.143-61, North Holland.

[17] Chamberlain, G. and Griliches, Z. (1975) "Unobservables with a Variance-Components Structure: Ability, Schooling and the Economic Success of Brothers," *International Economic Review*, 16, p.422-50.

[19] Cameron, A.C.and Trivedi, P.K.(1998) *Regression Analysis of Count Data*, Cambridge University Press.

[19] Cameron, A.C. and Trivedi, P.K.(2005) *Microeconometrics: Methods and Applications*, Cambridge University Press.

[20] Chao, John.C. and Swanson, Norman R.(2005) "Consistent Estimation with a Large Number of Weak Instruments", *Econometrica*, 73(5), PP.1673-1692.

[21] Cragg, John G.and Donald, Stephen G.(1993) "Testing Identifiability and Specification in Instrumental Varaible Models", *Econometric Theory*, 9, pp.222-40.

[22] Davidson, Russell and MacKinnon, James G.(2004) *Econometric Theory and Methods*, Oxford University Press.

[23] Durbin, J.(1954) "Errors in variables", *Review of the Internatinal Statistical Institute*, 22, pp.23-32.

[24] Griliches, Zvi.(1976) "Wages of Very Young Men", *Journal of Political Economy*, 84(4. Part 2), pp. S69-S85.

[25] Griliches, Zvi.(1977) "Estimating the Returns to Schooling: Some Econometric Problems", *Econometrica*, 45(1), pp.1-22.

[26] Griliches, Zvi., Hall, Bronwyn., and Hausoman, Jerry.(1978) "Missing Data and self-Selection in Large Panels", *Annales de L'INSEE*, XXX-XXXI, pp.137-76.

[27] Hahn, Jinyoung and Hausman, Jerry. (2002a) "A New Specification Test for the Validity of Instrumental variables", *Econometrica*, 70(1), pp.163-189.

[28] Hahn, Jinyoung and Hausman, Jerry. (2002b) "Notes on Bias in Estimators for Simultaneous Equation Models", *Economics Letters*, 75. pp.237-241.

[29] Hahn, Jinyoung and Hausman, Jerry. (2003) "Weak Instruments: Diagnosis and Cures in Empirical Econometrics", *American Economic Review*, 93(2), pp.118-125.

[30] Hall, Alastair R., Rudebusch, Glenn D. and Wilcox, David W.(1996) "JUdging Instrument Relevance in Instrumental Variables Estimation", *International Economic Review*, 37(2), pp.283-298.

[31] Hall, Alastair R. and Peixe, Fernanda P.M.(2000) "A Consistent Method for the Selection of Relevant Instruments", A paper presented at Econometric Society World Congress

2000.

[32] Hansen, Lars.P (1982) "Large Sample Properties of Generalized Method of Moments Estimators", *Econometrica*, 50(4), pp.1029-1054.

[33] Hausman, Jerry. (1978) "Specification tests in econometrics", *Econometrica*, 46, pp.1251-72.

[34] Hausman, Jerry., Stock, James H. and Yogo, Motohiro.(2005) "Asymptotic Properties of the Hahn-Hausman Test for Weak-Istruments", *Economics Letters*, 89, pp.333-42.

[35] Hausman, J.A. and Taylor, W.E. (1981) "Panel Data and Unobservable Individual Effects," *Econometrica*, 49, pp.1377-1398.

[36] Hayashi, Fumio.(2000) *Econometrics*, Princeton University Press.

[37] Hsiao, C.(1986) *Analysis of Panel Data*, Cambridge University Press.

[38] Hsiao, C. (2003) *Analysis of Panel Data 2nd ed.*, Cambridge University Press.

[39] Koenker, Roger.(1981) "A Note on Studentizing a test for Heteroscedasticity", *Journal of Econometrics*, 17., pp.107-112.

[40] Nelson, Charles R. and Startz, Richard.(1990a) "The Distribution of the Instrumental Variables Estimator and Its t-Ratio When the Instrument is a Poor One", *Journal of Business*, 63(1, Part.2), pp. S125-S140.

[41] Nelson, Charles R.and Startz, Richard.(1990) "Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimator", *Econometrica*, 58(4), pp.967-76.

[42] Pagan, A.R. and Hall, D. (1983) "Diagnostic Tests as Residual Analysis", *Econometric Reviews*, 2(2), pp.159-218.

[43] Ruud, P.A. (2000) *An Introduction to Classical Econometric Theory*, Oxford University Press.

[44] Sargan, J.D. (1958) "The Estimation of Economic Relationships Using Instrumental Variables", *Econometrica*, 26(3), pp.393-415.

[45] Shea, John.(1997) "Instrument Relevance in Multivariate Linear Models: A Simple Measure", *Review of Economics and Statistics*, 79(2), pp.348-352.

[46] Staiger, Douglas. and Stock, James.H. (1997) "Instrumental Variables Regression with Weak Instrumetns", *Econometrica*, 65(3), pp.557-86.

[47] Stock, James H. and Wright Jonathan H. (2000) "GMM with Weak Identification", *Econometrica*, 68(5), pp.1055-96.

[48] Stock, James H., Wright, Jonathan H. and Yogo, Motohiro. (2002) "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments", *Journal of Business and Economic Statistics*, 20(4), pp.518-29.

[49] Stock, James H. and Yogo, Motohiro. (2005) "Testing for Weak Instruments in Linear IV Regression", in Andrews, D.W.K. and Stock, J.H.(eds) *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, Cambridge University

Press. pp.80-108.

[50] Winklemann, Rainer and Boes, Stefan. (2005) *Analysis of Microdata*, Springer.

[51] White, Halbert. (1980) "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity", *Econometrica*, 48(4), pp.817-838.

[52] White, Halbert. (1982) "Instrumental Variables Regression with Independent Observations", *Econometrica*, 50(2), pp.483-499.

[53] Wooldridge, Jeffrey. M. (2002) *Econometric Analysis of Cross Section and Panel Data*, The MIT Press

[54] Wu, D-M. (1973) "Alternative tests of independence between stochastic regressors and disturbances", *Econometrica*, 41, pp.733-50.

第 4 章

生存時間分析

4.1 生存時間分析の基本量

パネルデータでは経時観察がされることから、特定のイベント発生までの長さを分析することに利用できるが、このような際に用いられるのが生存時間分析である。ここでは生存時間分析に使われる基本的な関数などについて簡単に説明する。[*1]

X をあるイベントが起きるまでの時間を表す確率変数であるとする。生存時間分析では、イベントの生起を死亡にみなして、イベントの起きるまでの時間を生存時間と呼ぶ。このとき、対象がある時間 x を越えて生存する確率は、

$$S(x) = Pr(X > x)$$

で定義されるが、これを生存関数という。生存関数は累積分布関数 $F(x) = Pr(X \leq x)$ と $S(x) = 1 - F(x)$ との関係にある。x が連続で確率密度関数 $f(x)$ が存在するとすれば、

$$S(x) = Pr(X > x) = \int_x^{\infty} f(t)dt$$

となるので、

$$f(x) = -\frac{dS(x)}{dx}$$

が成立する。さらに、

$$h(x) = \lim_{\Delta x \to 0} \frac{Pr(x \leq X < x + \Delta x | X \geq x)}{\Delta x}$$

をハザード関数と呼ぶ。連続な場合には、

$$h(x) = \frac{f(x)}{S(x)} = -\frac{d}{dx} \log S(x)$$

が成立する。

なお、生命表では、$S(x)$ を l_x、$f(x)$ を d_x、$h(x)$ を μ_x と表している。

[*1] 本章の内容については、Klein and Moeschberger (2009)、中澤 (2007) を参考にしている。

4.2 カプラン・マイヤ推定量

Freinreich et al による白血病治療データ (Gehan データと呼ぶ) は、急性白血病にかかり、寛解の状態にある 42 人の子供について、プラセボと抗がん剤の 6-MP(6-メルカプトプリン) を投与して再発までの時間（月数）を観測したデータである。データは両群に関し、対象に関する時間と、その時間がイベント発生時間か、打ち切り時間かが示されている。まず、データを見てみよう。

Gehan データの読み込みと表示

```
library(survival)
library(MASS)
data(gehan)
gehan
```

出力結果

```
   pair time cens   treat
1     1    1    1 control
2     1   10    1    6-MP
3     2   22    1 control
4     2    7    1    6-MP
5     3    3    1 control
6     3   32    0    6-MP
7     4   12    1 control
8     4   23    1    6-MP
9     5    8    1 control
10    5   22    1    6-MP
11    6   17    1 control
12    6    6    1    6-MP
13    7    2    1 control
14    7   16    1    6-MP
15    8   11    1 control
16    8   34    0    6-MP
17    9    8    1 control
18    9   32    0    6-MP
19   10   12    1 control
20   10   25    0    6-MP
21   11    2    1 control
22   11   11    0    6-MP
23   12    5    1 control
24   12   20    0    6-MP
25   13    4    1 control
26   13   19    0    6-MP
27   14   15    1 control
28   14    6    1    6-MP
29   15    8    1 control
30   15   17    0    6-MP
31   16   23    1 control
32   16   35    0    6-MP
33   17    5    1 control
34   17    6    1    6-MP
35   18   11    1 control
36   18   13    1    6-MP
37   19    4    1 control
38   19    9    0    6-MP
39   20    1    1 control
40   20    6    0    6-MP
41   21    8    1 control
42   21   10    0    6-MP
```

このうち、6-MP を投与したグループについて、時間でソートしたデータを表示してみる。

コードと出力結果

```
gehanT <- subset(gehan, treat == "6-MP")
gehanT[order(gehanT$time),]

   pair time cens treat
12    6    6    1  6-MP
28   14    6    1  6-MP
34   17    6    1  6-MP
40   20    6    0  6-MP
4     2    7    1  6-MP
38   19    9    0  6-MP
2     1   10    1  6-MP
42   21   10    0  6-MP
22   11   11    0  6-MP
36   18   13    1  6-MP
14    7   16    1  6-MP
30   15   17    0  6-MP
26   13   19    0  6-MP
24   12   20    0  6-MP
10    5   22    1  6-MP
8     4   23    1  6-MP
20   10   25    0  6-MP
6     3   32    0  6-MP
18    9   32    0  6-MP
16    8   34    0  6-MP
32   16   35    0  6-MP
```

ここで、イベントが起きた時刻 t_i において、イベントを体験する可能性のある個体数を Y_i、発生したイベントの数を d_i、打ち切りの数を c_i とする。このとき、以下のような表が作成できる。

t_i	Y_i	d_i	c_i
0	21		
6	21	3	1
7	17	1	0
9	16	0	1
10	15	1	1
11	13	0	1
13	12	1	0
16	11	1	0
17	10	0	1
19	9	0	1
20	8	0	1
22	7	1	0
23	6	1	0

そこで、d_i のある時刻だけに改めてインデックス i を振り直し、以下のような生存関数の推定量を考える。

$$\hat{S}(t) = \begin{cases} 1 & (t < t_1) \\ \prod_{i(t_i<t)} \left(1 - \frac{d_i}{Y_i}\right) & (t \geq t_1) \end{cases}$$

例えば、$0 \leq t \leq 6$ では、$\hat{S}(t) = 1$、$6 \leq t \leq 7$ では、$\hat{S}(t) = 1 - \frac{3}{21}$、$7 \leq t \leq 10$ では、$\hat{S}(t) = (1 - \frac{3}{21})(1 - \frac{1}{17})$ などとなる。これをカプラン・マイヤ推定量と呼ぶ。