7.5　gretl による回帰分析（系列相関，不均一分散，操作変数法の例）

gretl のインストール後，デスクトップに

のアイコンをクリックする。

●データ入力について：　Excel でデータ・ファイルを作り，gretl に読み込ませる。

12 月 10 日の授業の例（458 ページ，都道府県別データで消費関数の推定）を利用する。

次の Excel ファイルのファイル名を「pref.xlsx」として保存する。HP からダウンロード可。

gretl のデフォルトのフォルダ（Documents¥gretl）に保存しているものとする。

都道府県データで，『県民経済計算 2017 年版』から利用。

変数名リスト：

C_i ＝ 家計最終消費支出（1 兆円）

Y_i ＝ 県民可処分所得（1 兆円）

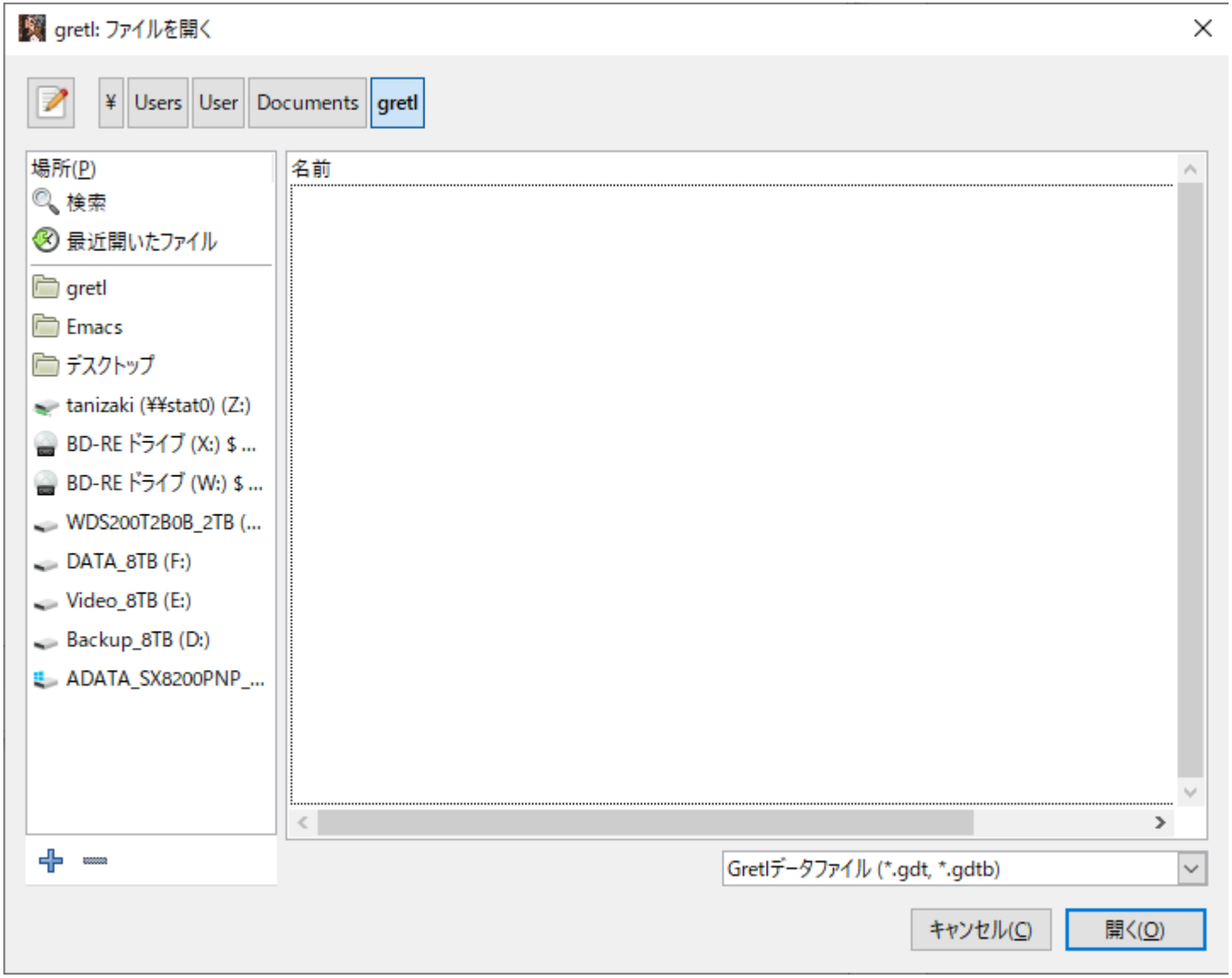
L_i ＝ 人口（千万人）

$i = 1, 2, \cdots, n$

$n = 47$

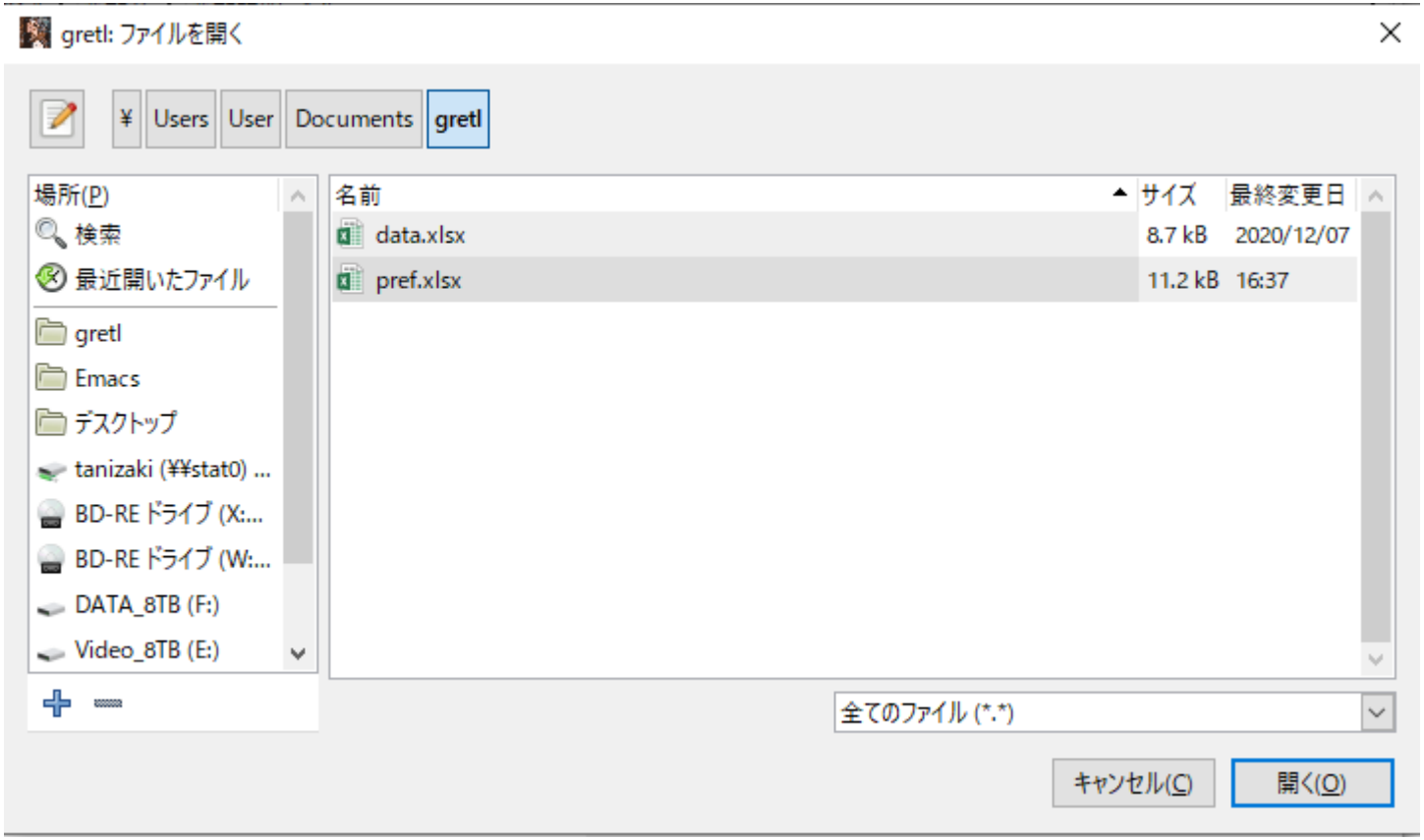
	A	B	C	D
1		c	yd	l
2	北海道	11.80617	18.79186	0.53201
3	青森県	2.71829	4.23312	0.12785
4	岩手県	2.73302	4.54180	0.12548
5	宮城県	5.04601	8.04722	0.23233
6	秋田県	2.11920	3.74344	0.09956
7	山形県	2.39465	4.01391	0.11017
8	福島県	3.98603	7.56076	0.18823
9	茨城県	6.20708	11.49900	0.28922
10	栃木県	4.43562	7.94783	0.19569
11	群馬県	4.25849	7.69941	0.19598
12	埼玉県	17.68357	24.79982	0.73096
13	千葉県	15.50630	22.34454	0.62456
14	東京都	43.65252	75.19110	1.37238
15	神奈川県	23.57387	33.81124	0.91587
16	新潟県	5.17907	8.37503	0.22665
17	富山県	2.44825	4.65801	0.10560
18	石川県	2.62807	4.20716	0.11475
19	福井県	1.79860	3.14056	0.07786
20	山梨県	1.80626	3.10818	0.08233
21	長野県	4.65047	7.76868	0.20758
22	岐阜県	4.21802	7.51484	0.20083
23	静岡県	8.54483	14.40090	0.36754
24	愛知県	18.61129	29.03244	0.75248
25	三重県	3.79440	6.64680	0.17996
26	滋賀県	3.13707	5.56586	0.14125
27	京都府	6.21440	9.50409	0.25992
28	大阪府	21.71001	30.72588	0.88233
29	兵庫県	12.75580	19.97605	0.55031
30	奈良県	3.12925	4.50147	0.13476
31	和歌山県	1.95272	3.31455	0.09449
32	鳥取県	1.14077	1.90643	0.05651
33	島根県	1.43286	2.68085	0.06849
34	岡山県	4.25354	6.70518	0.19071
35	広島県	6.44174	10.60937	0.28287
36	山口県	2.95729	6.08099	0.13829
37	徳島県	1.66517	2.74924	0.07433
38	香川県	2.23764	3.54480	0.09674
39	愛媛県	2.92152	4.62765	0.13641
40	高知県	1.49452	2.61836	0.07137
41	福岡県	11.06004	17.47809	0.51067
42	佐賀県	1.68834	2.89901	0.08238
43	長崎県	2.66037	4.03484	0.13540
44	熊本県	3.41067	6.20836	0.17653
45	大分県	2.36142	4.27603	0.11523
46	宮崎県	2.15512	3.57974	0.10888
47	鹿児島県	3.32918	4.99220	0.16257
48	沖縄県	2.65263	4.36765	0.14431

gretl で「ファイル」,「データを開く(O)」,「ユーザー・ファイル(U)」とし，次の画面になる。

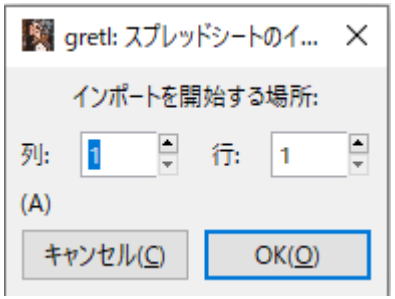


右下の「Gretl データファイル(*.gdt, *.gdtb)」のところを「全てのファイル(*.*)」にする。

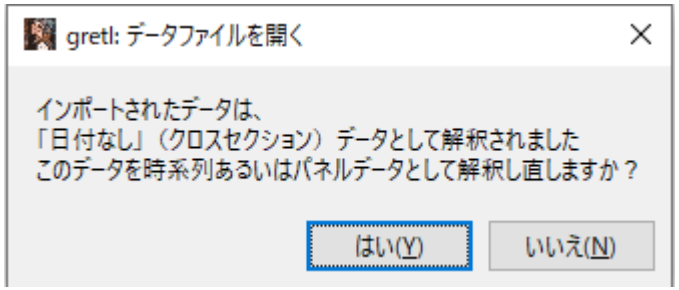
pref.xlsx ファイルが出てくる。



pref.xlsx を選択すると次の画面が出てくる。

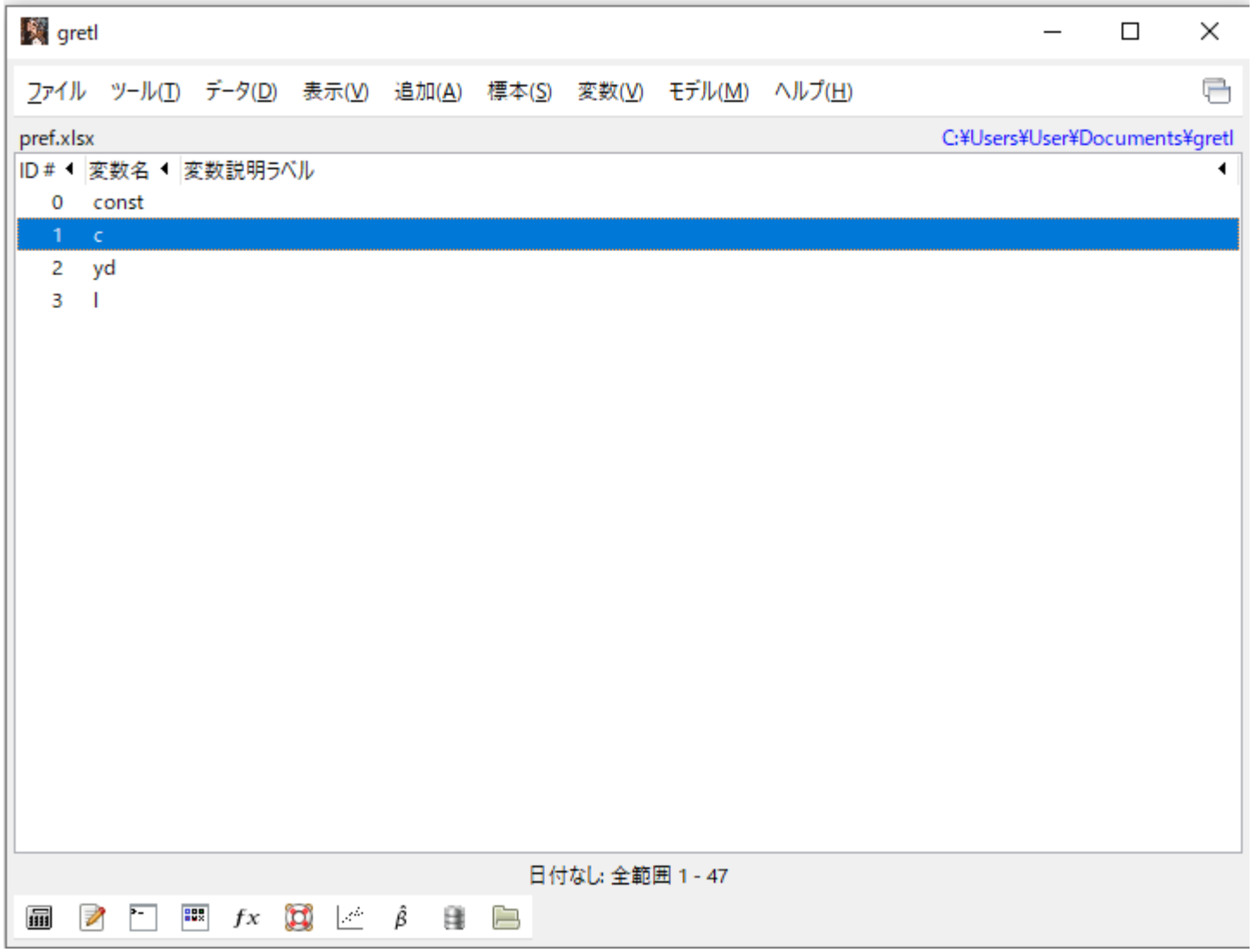


この場合は「OK(O)」で下の画面となる。



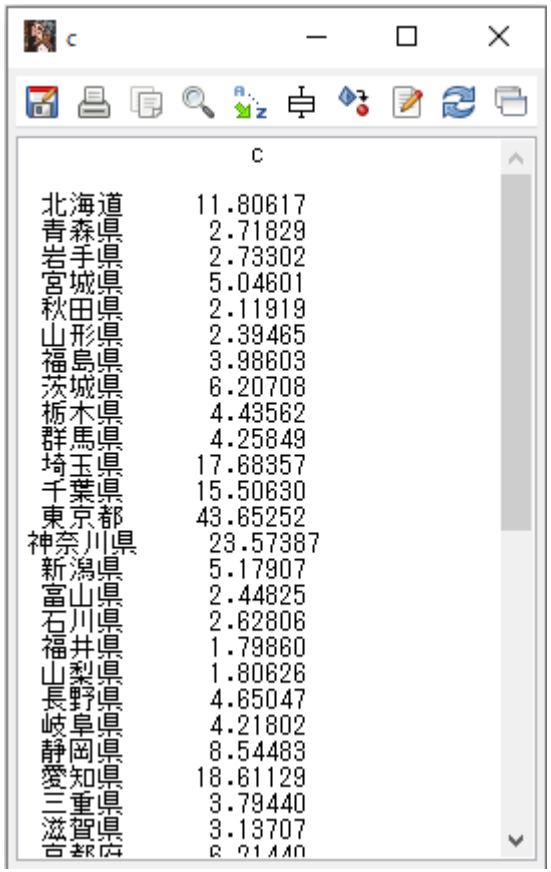
今回は，クロスセクション・データなので，「いいえ(N)」を選択する。

次の画面へ。



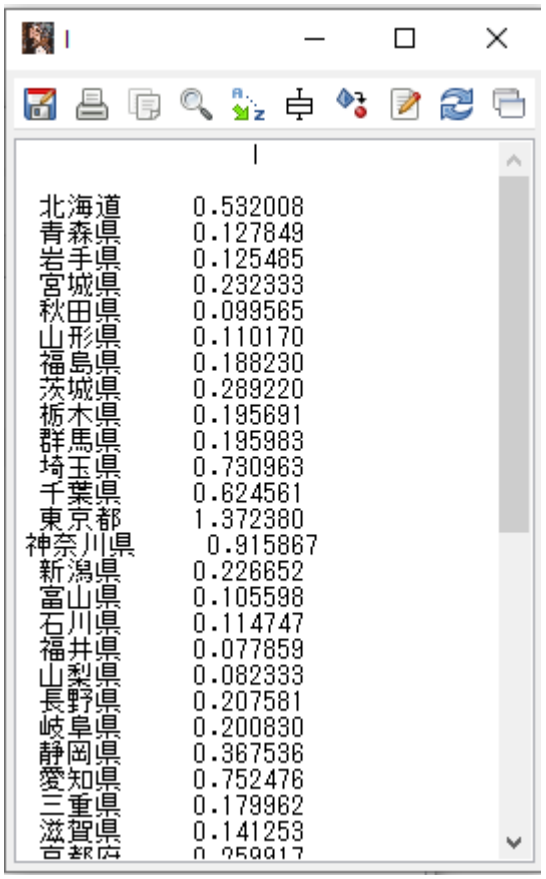
どのようにデータが入力されているかを確かめるために，順番に「c」，「yd」，「l」をクリックしていく。

「c」（消費）のデータ



「yd」（所得）のデータ

「l」（人口）のデータ



このデータを用いて，不均一分散の推定，説明変数と誤差項に相関がある場合に用いる二段階最小二乗法（two-stage least squares method）を解説する。

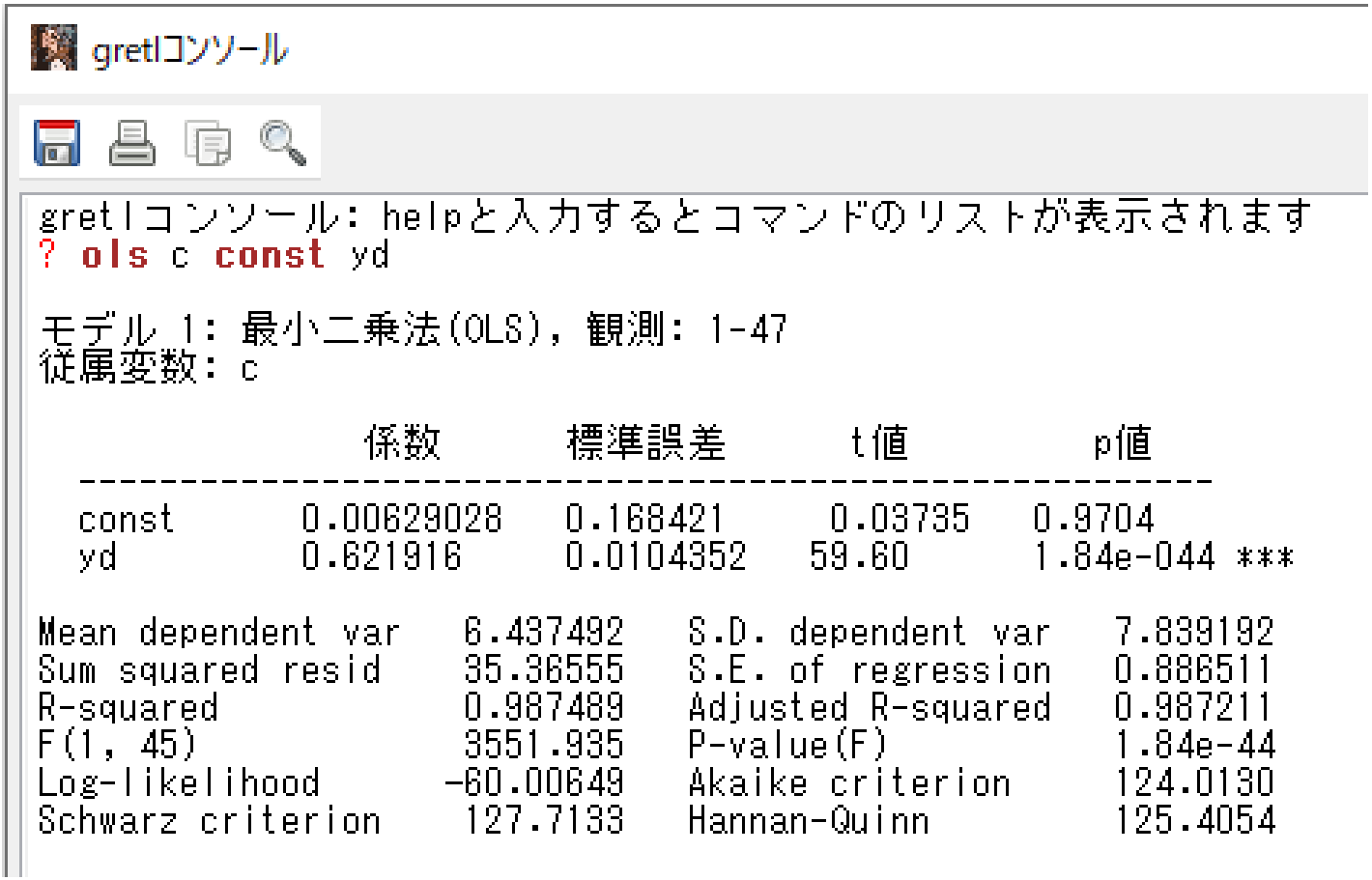
今回は前述の「● 推定方法　その２」の画面下の左から３番目の「  」(「gretl コンソールを開く」) を利用する。

●不均一分散について：

はじめに，$C_i = \alpha + \beta Yd_i + u_i$ を推定する。

「?」のあとに「ols c const yd」とタイプする（色は自動的につく）。

次の推定結果となる。



残差が人口（L_i）に依存するかどうかを検定する。直前の推定式からの残差は「$uhat」で表される。

残差の二乗は「$uhat^2」と表され，「u2」という変数を定義して，「u2」に「$uhat^2」を代入する。

「?」のあとに，「genr u2=$uhat^2」とタイプする。同様に，$L_i$の二乗を作るために，「genr l2=l^2」として，

「l2」を定義する。

このように，変数を定義した上で，$\hat{u}_i^2 = \alpha + \beta L_i^2 + u_i$ を推定するために「ols u2 const l2」とタイプする。

次の推定結果となる。

```
? genr u2=$uhat^2
系列 u2 (ID 4) を作成しました
? genr l2=l^2
系列 l2 (ID 5) を作成しました
? ols u2 const l2

モデル 2: 最小二乗法(OLS), 観測: 1-47
従属変数: u2

                係数       標準誤差      t値         p値
  ---------------------------------------------------------
  const       -0.0838041    0.131947    -0.6351    0.5286
  l2           5.69086      0.372018    15.30      1.91e-019 ***

Mean dependent var   0.752458   S.D. dependent var   2.027606
Sum squared resid    30.50170   S.E. of regression   0.823295
R-squared            0.838713   Adjusted R-squared   0.835129
F(1, 45)             234.0059   P-value(F)           1.91e-19
Log-likelihood      -56.52953   Akaike criterion     117.0591
Schwarz criterion    120.7593   Hannan-Quinn         118.4515
```

l2 の係数推定値の t 値は大きく，L_i^2が $\hat{u}_i^2$に影響していることは明らか。

すなわち，u_iの分散は不均一という結果になる。

したがって，最初の「ols c const yd」とタイプして得られた推定結果は，係数の標準誤差，t 値，p 値，回帰式の標準誤差などどの推定値も正しく推定されていないことになる。

そのため，これらの推定結果を使って，係数の信頼区間，仮説検定などは正しくできない。

正しい推論ができるようにするために，定数項も含めて，C_i，Yd_i を L_i で割って，推定し直す。

$C_i^*=C_i/L_i$ を「cl」，$Yd_i^*=Yd_i/L_i$ を「ydl」，$L_i^*=1/L_i$ を「l1」として変数を，「genr コマンド」を使って定義しなおす。

$C_i^*=\alpha L_i^*+\beta Yd_i^*+u_i^*$ を推定するために，「ols cl l1 ydl」とタイプする（「const」を含めない）。

```
? genr cl=c/l
系列 cl (ID 6) を作成しました
? genr ydl=yd/l
系列 ydl (ID 7) を作成しました
? genr l1=1/l
系列 l1 (ID 8) を作成しました
? ols cl l1 ydl

モデル 4: 最小二乗法(OLS), 観測: 1-47
従属変数: cl

              係数       標準誤差      t値        p値
  ---------------------------------------------------
  l1         -0.123386   0.0647118   -1.907   0.0630    *
  ydl         0.620603   0.0136731   45.39    3.17e-039 ***

Mean dependent var    22.20299   S.D. dependent var    2.086172
Sum squared resid     156.8788   S.E. of regression    1.867136
Uncentered R-squared  0.993287   Centered R-squared    0.216379
F(2, 45)              3329.279   P-value(F)            1.27e-49
Log-likelihood       -95.01527   Akaike criterion      194.0305
Schwarz criterion     197.7308   Hannan-Quinn          195.4230
```

「l1」の係数がもともとの回帰式の定数項α,「ydl」の係数が傾きβに相当する。

上記結果を用いて，α，βの信頼区間，仮説検定が正しくできるようになる。

一般的に，もともとの回帰式の傾きβの推定値の標準誤差は大きくなる傾向が強い（0.0104352 から 0.0136731 へ）。

● 二段階最小二乗法（操作変数法の一つ）について：

実は，$C_i = \alpha + \beta Yd_i + u_i$ の推定は問題があって，説明変数 Yd_i と誤差項 u_i には相関がある。

説明変数と誤差項に相関がある場合，α，β の推定量は不偏推定量にも一致推定量にもならない。

まず，説明変数 Yd_i と誤差項 u_i には相関があることを示す。

「可処分所得（Yd_i）＝消費（C_i）＋貯蓄（S_i）」が成り立つ。

(*) 可処分所得とは，労働の対価として得た給与やボーナスなどの個人所得から，支払い義務のある税金や社会保険料などを差し引いた，残りの手取り収入のこと。個人が自由に使用できる所得の総額。

Yd_i に C_i が含まれるため，Yd_i は u_i に依存することになる。

したがって，説明変数 Yd_i と誤差項 u_i には相関がある。

$C_i = \alpha + \beta Yd_i + u_i$ を正しく推定するためには（α, β の一致推定量を得るためには），操作変数法を用いる必要がある。

代表的な操作変数法として，二段階最小二乗法がよく用いられる。

$Z_i = \hat{Yd}_i$（Yd_i の予測値）を操作変数として用いる。

$Yd_i = \gamma_0 + \gamma_1 L_i + v_i$ を最小二乗法で推定して，Yd_i の予測値を求める。

以上を，gretl で表すと，「tsls c const yd; const l」とタイプする。

tsls ＝ 二段階最小二乗法（two-stage least squares method）

セミコロン「;」以下の変数は，「;」以下の変数を用いて説明変数の予測値を求めて，その予測値を操作変数として用いるという意味である。

結果は次の通り。

```
? tsls c const yd; const |

モデル 8: 二段階最小二乗法(2SLS), 観測: 1-47
従属変数: c
内生変数(instrumented): yd
操作変数: const |

              係数       標準誤差       t値         p値
  ------------------------------------------------------------
  const      -0.159527   0.175057    -0.9113   0.3670
  yd          0.637951   0.0110454   57.76     7.41e-044 ***

Mean dependent var   6.437492   S.D. dependent var   7.839192
Sum squared resid    37.22125   S.E. of regression   0.909472
R-squared            0.987489   Adjusted R-squared   0.987211
F(1, 45)             3335.911   P-value(F)           7.41e-44
Log-likelihood      -230.4391   Akaike criterion     464.8782
Schwarz criterion    468.5785   Hannan-Quinn         466.2706

ハウスマン(Hausman)検定 -
  帰無仮説: OLS推定値は一致性を持つ
  漸近的検定統計量: カイ二乗(1) = 204.808
  なお、p値(p-value) = 1.86481e-046

弱操作変数(weak instrument)の検定 -
  第1段階のF統計量 (1, 45) = 697.529
  名目5%の有意水準で検定を行う場合の望ましいTSLS最大サイズに対する臨界値:

     size      10%      15%      20%      25%
    value    16.38     8.96     6.66     5.53
```

モデル 3: 二段階最小二乗法(2SLS), 観測: 1-47
従属変数: c
内生変数(instrumented): yd
操作変数: const I

	係数	標準誤差	t 値	p 値	
const	−0.159527	0.175057	−0.9113	0.3670	
yd	0.637951	0.0110454	57.76	7.41e-044	***

基本的には, 推定結果の上記の部分がわかっていれば, 全く問題ない。

「従属変数」＝「被説明変数」

「内生変数」＝「説明変数の中で誤差項と相関のある変数」

「操作変数」＝この場合は,「const, I を説明変数として, yd の予測値を求める」という意味

係数, 標準誤差, t 値, p 値は今までと同じ。

推定　その２（不均一分散＋二段階最小二乗法）:

$C_i = \alpha + \beta Yd_i + u_i$ について，u_i の分散は L_i に依存して不均一という結果が得られたので，データを変換して，$C_i^* = \alpha L_i^* + \beta Yd_i^* + u_i^*$ を二段階最小二乗法で推定する。

それぞれの記号は，$C_i^* = C_i/L_i$，$L_i^* = 1/L_i$，$Yd_i^* = Yd_i/L_i$，$u_i^* = u_i/L_i$ である。

変換した変数に関しても，Yd_i^* と u_i^* は相関がある。

(*) なぜなら，Yd_i^* は Yd_i の関数で，Yd_i は C_i に依存し，C_i は u_i に依存する。

結果として，Yd_i^* は u_i の関数となっている。

u_i^* は u_i の関数であり，Yd_i^* と u_i^* はともに同じ確率変数 u_i を含んでいる。

したがって，Yd_i^* と u_i^* は相関がある。

$C_i^* = C_i/L_i$ を「cl」，$Yd_i^* = Yd_i/L_i$ を「ydl」，$L_i^* = 1/L_i$ を「l1」とそれぞれ変数を定義しているので，

「tsls cl l1 ydl; const l1 l」

として，推定結果が得られる。

```
? tsls cl l1 ydl; const l1 l

モデル 2: 二段階最小二乗法(2SLS), 観測: 1-47
従属変数: cl
内生変数(instrumented): ydl
操作変数: const l1 l

              係数       標準誤差      t値         p値
  ------------------------------------------------------------
  l1         -0.179116   0.0662284   -2.705    0.0096    ***
  ydl         0.634558   0.0140590   45.14     4.05e-039 ***

Mean dependent var    22.20299   S.D. dependent var    2.086172
Sum squared resid     160.5102   S.E. of regression    1.888622
Uncentered R-squared  0.529569   Centered R-squared    0.638500
カイ二乗(2)              6531.605   p値                    0.000000

ハウスマン(Hausman)検定 -
  帰無仮説: OLS推定値は一致性を持つ
  漸近的検定統計量: カイ二乗(1) = 106.962
  なお、p値(p-value) = 4.53744e-025

Sarganの過剰識別検定 -
  帰無仮説: 全ての操作変数は有効(valid)である
  検定統計量: LM = 1.59516
  なお、p値(p-value) = P(カイ二乗(1) > 1.59516) = 0.20659

弱操作変数(weak instrument)の検定 -
  第1段階のF統計量 (2, 44) = 660.273
  名目5%の有意水準で検定を行う場合の望ましいTSLS最大サイズに対する臨界値:

サイズ      10%      15%      20%      25%
      値    19.93    11.59     8.75     7.25
```

モデル 2: 二段階最小二乗法(2SLS), 観測: 1-47
従属変数: cl
内生変数(instrumented): ydl
操作変数: const l1 l

	係数	標準誤差	t 値	p 値	
l1	−0.179116	0.0662284	−2.705	0.0096	***
ydl	0.634558	0.0140590	45.14	4.05e-039	***

この場合も同様に, 推定結果の上記の部分がわかっていれば, 全く問題ない。

(*) 二つの説明変数 l1, ydl をそれぞれ const, l1, l で回帰させて, それぞれの予測値を操作変数として用いるというものである。

すなわち,

・$L_i^* = \gamma_0 + \gamma_1 L_i^* + \gamma_2 L_i + v_i$ を最小二乗法で推定し, L_i^*の予測値 $\hat{L}_i^*$を求める（この場合, $\hat{L}_i^* = L_i^*$)。

・$Yd_i^* = \delta_0 + \delta_1 L_i^* + \delta_2 L_i^* + w_i$ を最小二乗法で推定し, Yd_i^*の予測値 $\hat{Yd}_i^*$を求める。

となる。

$\hat{L}_i^* = L_i^*$の理由：

最小二乗法によると残差平方和を最小にするようなパラメータの推定値が得られる。

$\gamma_0=0$, $\gamma_1=1$, $\gamma_2=0$ とすれば，残差はゼロになる（残差平方和もゼロ）。

(*) 説明変数にある変数を操作変数に使う場合は，その説明変数は誤差項とは相関がないと仮定していることに等しい。