第２回：統計解析ソフトへのデータの取り込み

北村　友宏

2020 年 5 月 15 日

本日の内容

1. ダミー変数
2. gretl でのデータの取り込み
3. gretl での変数の作成
4. gretl での記述統計の出力

ダミー変数

- ある事柄が当てはまるなら 1，当てはまらないなら 0 とする変数をダミー変数（dummy variable）という.
- 前回の実習では，「男性」と「女性」の 2 つのダミー変数を作成した.
  - 変数「男性」は，男性なら 1，女性なら 0 とした.
  - 変数「女性」は，女性なら 1，男性なら 0 とした.

加工・整理後の Excel ファイルの形

	A	B	C	D	E	F
1	id	prefecture	income	consumption	male	female
2	1	Hokkaido	227,349	155,491	1	0
3	2	Aomori	233,967	175,207	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
48	47	Okinawa	214,233	137,726	1	0
49	1	Hokkaido	207,155	172,835	0	1
50	2	Aomori	169,422	143,179	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
95	47	Okinawa	144,644	139,716	0	1

統計解析ソフト gretl

- 統計解析ソフト gretl は，無料でダウンロード・インストール・利用できる．
- Excel ファイルや csv ファイルのデータセットを取り込むことができる．
  - Excel ファイルについては，現行バージョンであれば xls，xlsx 両方に対応．
- 現行バージョンは日本語に対応．
- マウス操作で分析を実行する．

実習 1

最新バージョンの統計解析ソフト gretl を入手し，自分の PC にインストールする．
※すでに自分の PC に gretl をインストールしていても，2020 年 4 月 10 日以前にダウンロード・インストールした場合は再度，最新バージョンをダウンロードし，再インストールすること．

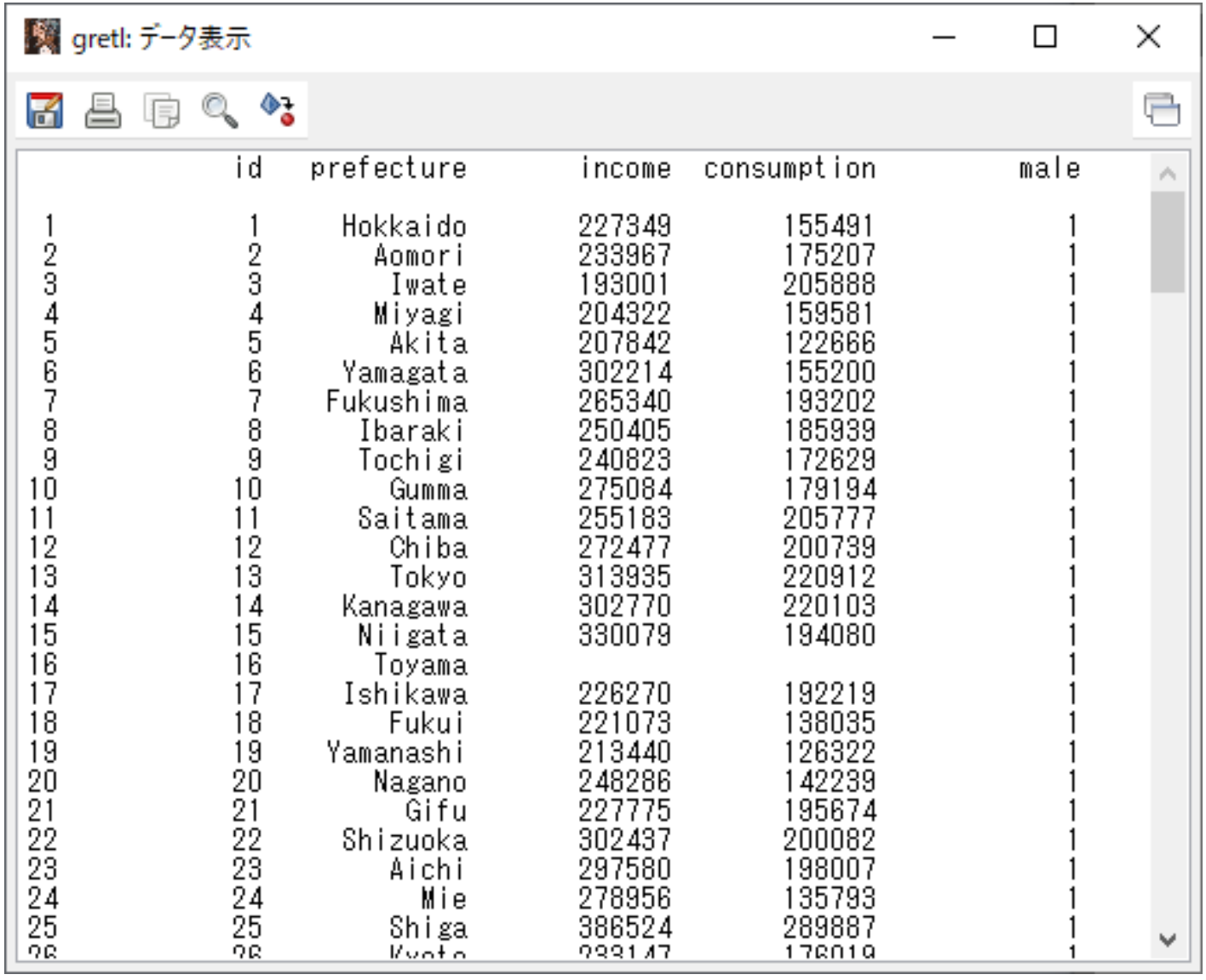
1. gretl の公式 HP（http://gretl.sourceforge.net/）にアクセス．
2. Windows の場合は「gretl for Windows」を，Mac の場合は「gretl on Mac OS X」をクリック．

3. latest release にあるリンクをクリックしてインストールファイルを保存.
   - Windows の場合：
     最近の PC はほとんど 64bit 版なので,
     gretl-2020b-64.exe を選んでも問題ない場合が多い. 自分の PC が 32bit 版であれば,
     gretl-2020b.exe を選ぶ. 解凍ソフト（7-Zip や Lhaplus など）を持っていれば,
     gretl-2020b-win32.zip を選んでもよい.
   - Mac の場合：
     gretl-2020b-quartz.pkg を選ぶ.
4. 保存したインストールファイルを実行してインストールまたは解凍.

実習 2

1. 先ほどの実習でインストールした gretl を起動.
2. 前回の実習で作成した消費 2009.xlsx を，gretl の画面にドラッグ・アンド・ドロップ.
3. 出てきたダイアログボックスの，インポートを開始する場所: の列: と行: がともに 1 になっていることを確認し，「OK」をクリック.
4. 「インポート可能なシートを 1 個見つけました」で始まるメッセージが表示されるので，「閉じる」をクリックすると，データが読み込まれる.

5. 「インポートされたデータは・・・(中略)・・・解釈し直しますか？」というメッセージが表示されるので，「いいえ」をクリックすると，データが読み込まれる．
   - 前回みたように，『全国消費実態調査』2009 年版の都道府県別・男女別の所得と消費のデータは「横断面（クロスセクション）データ」なので，「いいえ」でよい．時系列データやパネルデータを読み込む場合，このメッセージが表示されたら「はい」をクリックする．
6. 「id」から「female」までの 6 つをドラッグして選択し，その上で右クリック→「データ（値）を表示」と操作すると，全変数の観測値リストが新規ウィンドウにて表示される．



	id	prefecture	income	consumption	male
1	1	Hokkaido	227349	155491	1
2	2	Aomori	233967	175207	1
3	3	Iwate	193001	205888	1
4	4	Miyagi	204322	159581	1
5	5	Akita	207842	122666	1
6	6	Yamagata	302214	155200	1
7	7	Fukushima	265340	193202	1
8	8	Ibaraki	250405	185939	1
9	9	Tochigi	240823	172629	1
10	10	Gumma	275084	179194	1
11	11	Saitama	255183	205777	1
12	12	Chiba	272477	200739	1
13	13	Tokyo	313935	220912	1
14	14	Kanagawa	302770	220103	1
15	15	Niigata	330079	194080	1
16	16	Toyama			1
17	17	Ishikawa	226270	192219	1
18	18	Fukui	221073	138035	1
19	19	Yamanashi	213440	126322	1
20	20	Nagano	248286	142239	1
21	21	Gifu	227775	195674	1
22	22	Shizuoka	302437	200082	1
23	23	Aichi	297580	198007	1
24	24	Mie	278956	135793	1
25	25	Shiga	386524	289887	1
26	26	Kyoto	233147	176019	1

このような画面が表示されれば成功．確認したら閉じる．

※もし数字が違っていたら，データセット（消費2009.xlsx）の作成の際にミスをしているということなので，前回の講義スライドを参照してデータセットの作成からやり直すこと.

7. メニューバーから「ファイル」→「データに名前を付けて保存」と操作し，消費 2009.gdt という名前で「2020 ミクロデータ分析 1」フォルダに保存（全角日本語使用可）.

変数の作成方法

メニューバーから「追加」→「新規変数の定義」と操作し，入力ボックスに変数の定義式を入力する．

- 「ヘルプ」をクリックすると，演算子や関数の入力方法を参照できる．

変数の単位の変換

元のデータの可処分所得と消費支出は円単位.

⇓

そのままでは桁数が多く，出力結果が見にくい場合がある.

⇓

例えば千円単位にすると，出力結果が見やすくなる.

⇓

千円単位にするには，新たな変数を作成し，元の変数を 1,000 で割ったものと定義すればよい.

実習 3

1. gretl のメニューバーから「追加」→「新規変数の定義」と操作.
2. 入力ボックスに,

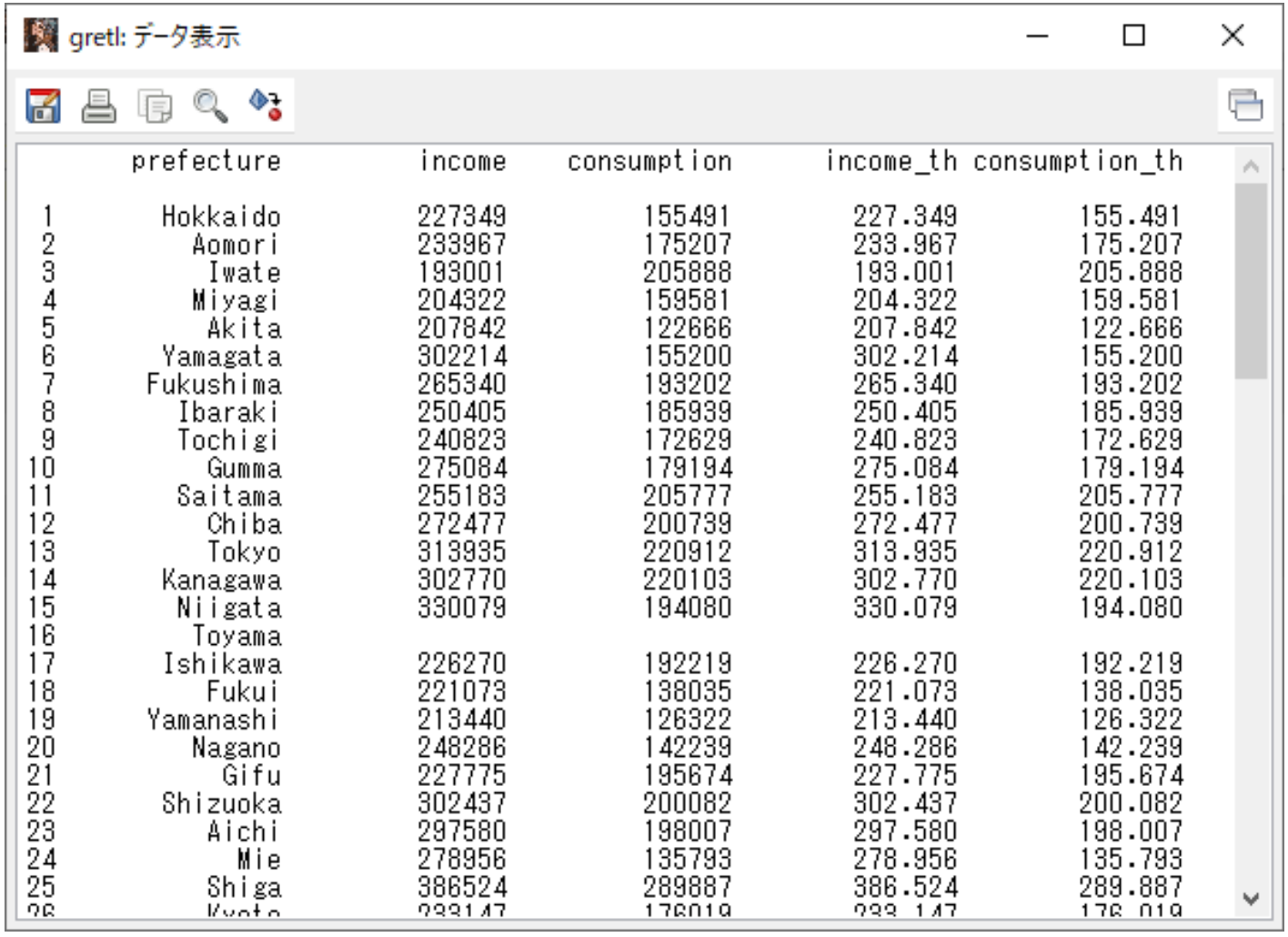
   income_th=income/1000

   と入力して「OK」をクリックすると,「変数 income の全観測値について 1000 で割ったもの」という定義の「変数 income_th」が作成される.
   - 左辺の変数名はなんでもよい. ここでは thousand（千）の th を付けて, 元の変数 income と区別する.
   - 割り算の演算子は「/（スラッシュ）」
3. gretl のメニューバーから「追加」→「新規変数の定義」と操作.

4. 入力ボックスに，

   consumption_th=consumption/1000

   と入力して「OK」をクリックすると，「変数 income の全観測値について 1000 で割ったもの」という定義の「変数 income_th」が作成される．
5. gretl のメニューバーから「ファイル」→「データを保存」と操作して上書き保存．
6. Ctrl キーを押しながら「prefecture」，「income」，「consumption」，「income_th」，「consumption_th」の 5 つをクリックして選択し，その上で右クリック→「データ（値）を表示」と操作すると，これら 5 つの変数の観測値リストが新規ウィンドウにて表示される．



	prefecture	income	consumption	income_th	consumption_th
1	Hokkaido	227349	155491	227.349	155.491
2	Aomori	233967	175207	233.967	175.207
3	Iwate	193001	205888	193.001	205.888
4	Miyagi	204322	159581	204.322	159.581
5	Akita	207842	122666	207.842	122.666
6	Yamagata	302214	155200	302.214	155.200
7	Fukushima	265340	193202	265.340	193.202
8	Ibaraki	250405	185939	250.405	185.939
9	Tochigi	240823	172629	240.823	172.629
10	Gumma	275084	179194	275.084	179.194
11	Saitama	255183	205777	255.183	205.777
12	Chiba	272477	200739	272.477	200.739
13	Tokyo	313935	220912	313.935	220.912
14	Kanagawa	302770	220103	302.770	220.103
15	Niigata	330079	194080	330.079	194.080
16	Toyama				
17	Ishikawa	226270	192219	226.270	192.219
18	Fukui	221073	138035	221.073	138.035
19	Yamanashi	213440	126322	213.440	126.322
20	Nagano	248286	142239	248.286	142.239
21	Gifu	227775	195674	227.775	195.674
22	Shizuoka	302437	200082	302.437	200.082
23	Aichi	297580	198007	297.580	198.007
24	Mie	278956	135793	278.956	135.793
25	Shiga	386524	289887	386.524	289.887
26	Kyoto	233147	176019	233.147	176.019

このような画面が表示されれば成功．確認したら閉じる．

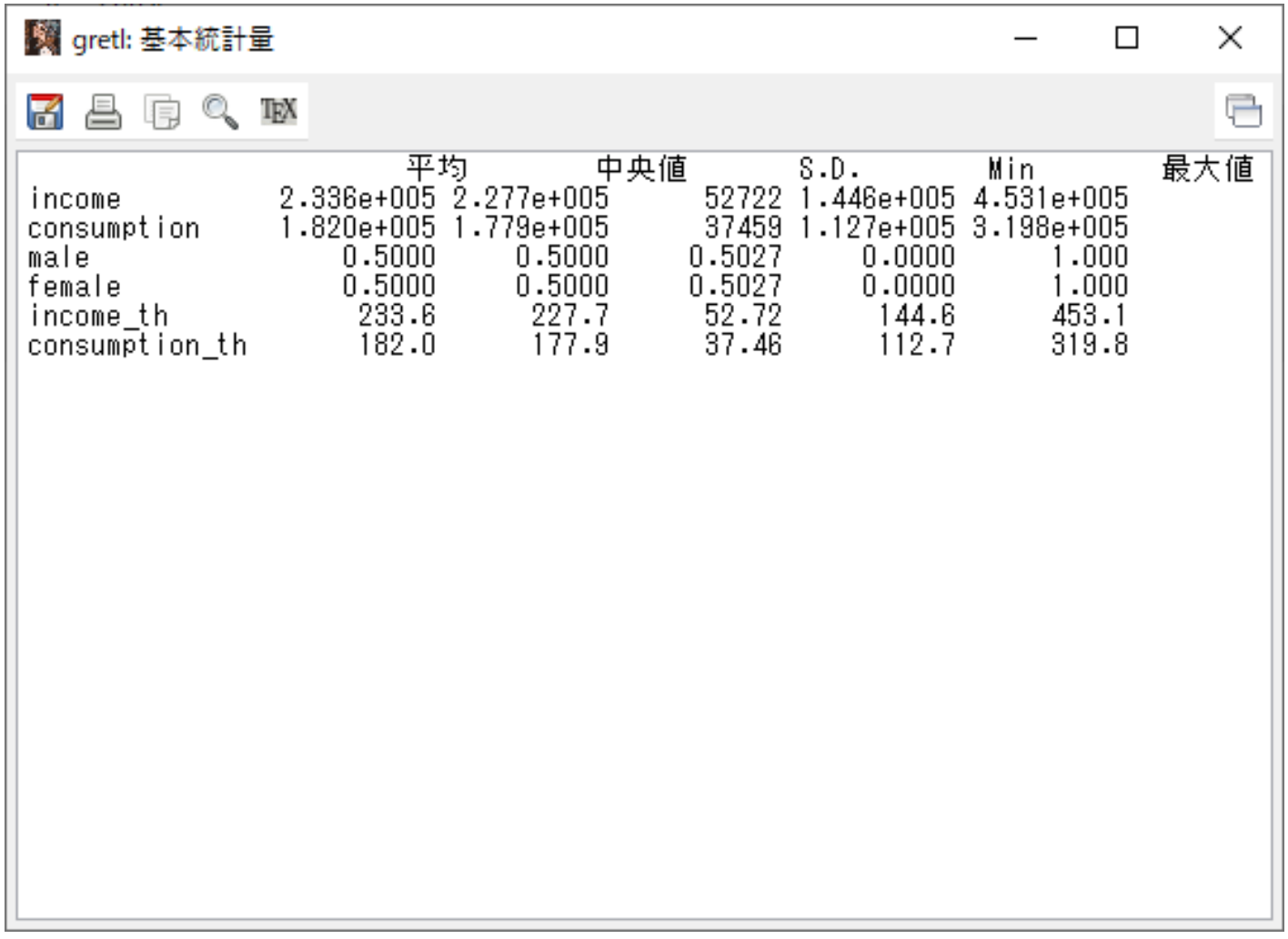
記述統計

- データセットを読み込んだ gretl の画面上で，記述統計を出力したい変数を選択し，右クリック→「基本統計量」と操作し，Show main statistics が選ばれている状態で「OK」をクリックすると，選んだ変数の，平均 (mean)，中央値 (median)，標準偏差 (standard deviation, S.D.)，最小値 (minimum, Min)，最大値 (maximum) が表示される．
  - 「記述統計」は，「基本統計量」や「要約統計量」ともいう．
  - 「Show full statistics」を選ぶことによって出力できる統計量は，次回の授業で解説する．

- 平均
  - $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$
- 中央値
  - 観測値を小さい順に並べたときに中央に来る値.
  - 観測値数 n が偶数の場合は中央で隣り合う 2 つの値の平均値.
- S.D.: 標準偏差
  - $s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}.$
- Min: 最小値
  - $\min\{x_i\}.$
- 最大値
  - $\max\{x_i\}.$

実習 4

1. 「income」から「consumption_th」までの 6 つをドラッグして選択し，その上で右クリック→「基本統計量」と操作.
2. Show main statistics が選ばれている状態で「OK」をクリックすると，選択した変数の記述統計 5 種類が表示される.



	平均	中央値	S.D.	Min	最大値
income	2.336e+005	2.277e+005	52722	1.446e+005	4.531e+005
consumption	1.820e+005	1.779e+005	37459	1.127e+005	3.198e+005
male	0.5000	0.5000	0.5027	0.0000	1.000
female	0.5000	0.5000	0.5027	0.0000	1.000
income_th	233.6	227.7	52.72	144.6	453.1
consumption_th	182.0	177.9	37.46	112.7	319.8

このような画面が表示されれば成功.
Mac の PC では，小数点以下の表示桁数が異なっている場合がある.

- 統計量の名前の位置がズレていて見づらいが，各変数について出力された数字は左から平均，中央値，標準偏差，最小値，最大値の順.

- e+005 は，$\times 10^5$ という意味.
  - e.g., 変数 income（円単位の可処分所得）の平均は 2.336×10^5（円）.

まだ作業があるので，「gretl: 基本統計量」のウィンドウはまだ閉じない！

3. 表示されている記述統計の画面上で右クリック→「名前を付けて保存...」と操作.
4. 出てきたダイアログボックスの,「標準テキスト」を選び,「OK」をクリック.
5. 記述統計 5 月 15 日.txt という名前で「2020 ミクロデータ分析 1」フォルダに保存. 本日の作業はここまで.