第 3 回：データの加工・整理（2）

北村　友宏

2020 年 10 月 16 日

本日の内容

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

実習 1

1. setagayaapartment.xlsx を開く.
2. F 列の,「1 R」と「1 K」となっているセルを全て半角の 1 に変更.
3. F 列の,「1 R」でも「1 K」でもなく，空白でもないセルを全て半角の 0 に変更．F 列の空白セルは空白のままにしておく.
   - 空白セルは計 4 個.
4. 上書き保存して閉じる.

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.
- 先ほどの実習では，onekr というダミー変数を作成した.
  - 部屋の種類が 1R または 1K なら 1，それ以外なら 0 とした.

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 2

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 8 月 6 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．

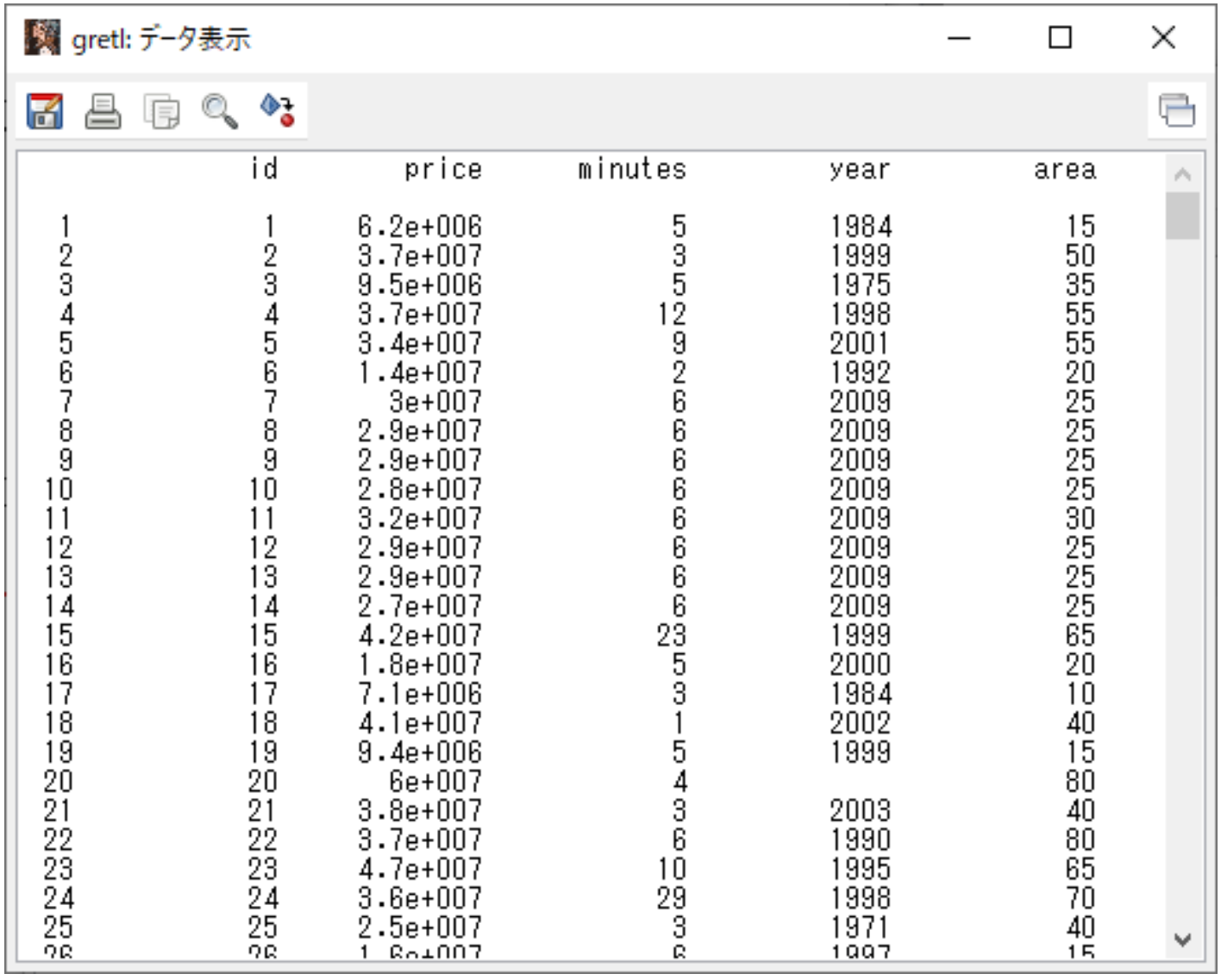
1. gretl の公式 HP（`http://gretl.sourceforge.net/`）にアクセス．
2. Windows の場合は「gretl for Windows」を，Mac の場合は「gretl on macOS」をクリック．

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-2020d-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-2020d-32.exe を選ぶ．解凍ソフト（7-Zip や Lhaplus など）を持っていれば,
     gretl-2020d-win32.zip を選んでもよい.
   - Mac の場合：
     gretl-2020d-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 3

1. 先ほどの実習でインストールした gretl を起動.
2. setagayaapartment.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．
   - 第１回の授業でみたように国土交通省　土地総合情報システム『不動産取引価格情報検索』の物件別のマンション価格などのデータは「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．
5. 「id」から「onekr」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	price	minutes	year	area
1	1	6.2e+006	5	1984	15
2	2	3.7e+007	3	1999	50
3	3	9.5e+006	5	1975	35
4	4	3.7e+007	12	1998	55
5	5	3.4e+007	9	2001	55
6	6	1.4e+007	2	1992	20
7	7	3e+007	6	2009	25
8	8	2.9e+007	6	2009	25
9	9	2.9e+007	6	2009	25
10	10	2.8e+007	6	2009	25
11	11	3.2e+007	6	2009	30
12	12	2.9e+007	6	2009	25
13	13	2.9e+007	6	2009	25
14	14	2.7e+007	6	2009	25
15	15	4.2e+007	23	1999	65
16	16	1.8e+007	5	2000	20
17	17	7.1e+006	3	1984	10
18	18	4.1e+007	1	2002	40
19	19	9.4e+006	5	1999	15
20	20	6e+007	4		80
21	21	3.8e+007	3	2003	40
22	22	3.7e+007	6	1990	80
23	23	4.7e+007	10	1995	65
24	24	3.6e+007	29	1998	70
25	25	2.5e+007	3	1971	40
26	26	1.6e+007	6	1997	15

このような画面が表示されれば成功．onekr の観測値リストは，下のほうに表示されている．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（setagayaapartment.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.

6. 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，setagayaapartment.gdt という名前で 2020microdatag フォルダに保存.

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，「主要な統計量を表示する」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)，最小値 (minimum)，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．

- 平均
  - $\bar{x} = \dfrac{1}{n} \sum_{i=1}^{n} x_i.$
- 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.
- 標準偏差
  - $s_x = \sqrt{\dfrac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}.$
- 最小値
  - $\min\{x_i\}.$
- 最大値
  - $\max\{x_i\}.$

分散と標準偏差

x_i の（標本）分散（sample variance）は，

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 .$$

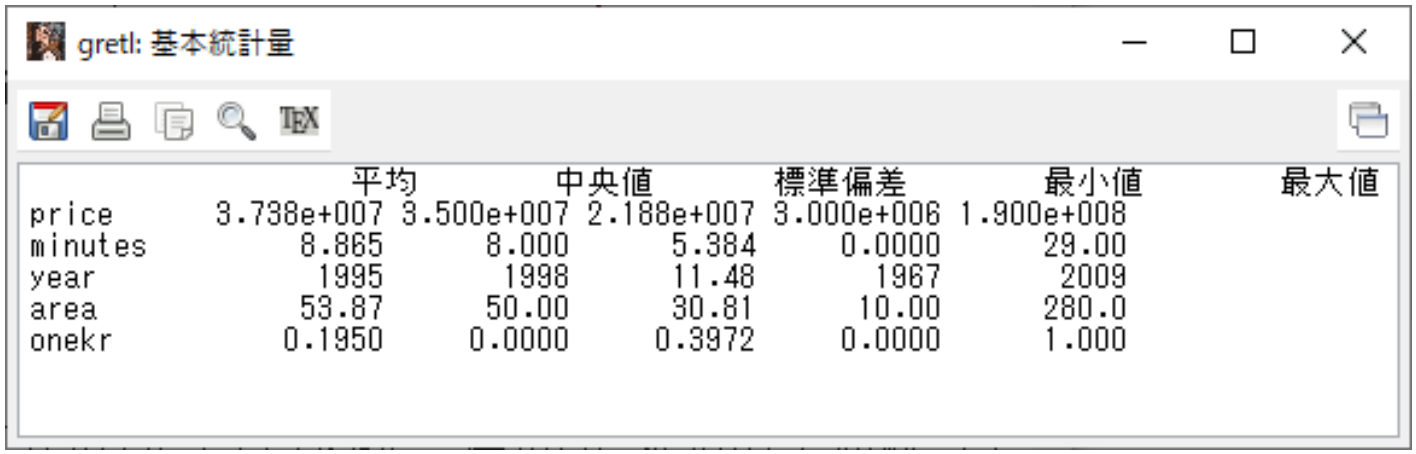
- x_i が個体間でどれだけバラついているかを測る．
- 単位が「x_i の単位の二乗」になる．
  - e.g., x_i の単位が「万円」$\Rightarrow$ s_x^2 の単位は「万円 ^{2}」

$\Rightarrow$ 正の平方根をとって（標本）標準偏差（sample standard deviation）に直すと，元の測定単位に戻る．

$$s_x = \sqrt{s_x^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

実習 4

1. 「price」から「onekr」までの 5 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. 「主要な統計量を表示する」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.
   - 最新バージョン（2020 年 8 月 6 日版）では，この表示が日本語化されている.



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.
最新バージョン（2020 年 8 月 6 日版）では，上の画像のように統計量名が全て日本語で表示される.

- 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+007 は，$\times 10^7$ という意味.
  - e.g., 変数 price（円単位の中古マンション価格）の平均は 3.738×10^7（円）.
- 「e-」の場合，例えば 2.5e-006 は，2.5×10^{-6} という意味.

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. summary20201016.txt という名前で 2020microdatag フォルダに保存. 本日の作業はここまで.