

# 第4回：データの加工・整理（3）

北村 友宏

2020年10月23日

# 本日の内容

1. gretl での観測値の消去
2. gretl での変数の作成

# 欠損値

- ▶ 現在分析しているデータの変数のうち、minutes, year, onekr には欠損値 (missing value) がある.
  - ▶ minutes: 4 個
  - ▶ year: 2 個
  - ▶ onekr: 4 個
- ▶ 欠損となっている変数をもつ個体は、(その変数を用いた) 線形回帰モデルを推定する際に無視される (使えない).



欠損となっている変数をもつ個体を、この段階で消去しておく。

# 欠損の観測値の消去方法

- ▶ gretl のメニューバーから「標本」→「欠損値を持つ観測を落とす」と操作し、OK をクリック。
  - ▶ すべての変数について、欠損値が1つでもあれば、その個体が消去される。
  - ▶ 「永続的に変更する」にチェックを**入れなければ**、消去後、**データセットを上書き保存していない状態**で、gretl のメニューバーから「標本」→「全範囲に戻す」と操作すると、観測値を消去する前のデータセットに戻すことができる。

# 実習 1

1. gretl を起動.
2. 「ファイル」 → 「データを開く」 → 「ユーザー・ファイル」と操作.
3. setagayaapartment.gdt を選択し, 「開く」をクリック.
4. gretl のメニューバーから「標本」 → 「欠損値を持つ観測を落とす」と操作.
5. OK をクリック.

6. 「10 個の観測を落としました」というメッセージが表示されるので、「閉じる」をクリック.
7. **上書き保存**する. メニューバーから「ファイル」→「データを保存」と操作すると、「データセットは現在、サブサンプルされています  
全範囲に戻しますか?」というメッセージが表示されるので、「**いいえ**」をクリック.

# gretl での変数の作成方法

1. gretl のメニューバーから「追加」→「新規変数の定義」と操作.
2. 出てきた「gretl: 変数の追加」ダイアログボックスの入力ボックスに

(付けたい変数名)=(変数の定義式)

を入力し、「OK」をクリック.

使える演算子などについては、「gretl: 変数の追加」ダイアログボックスの「ヘルプ」をクリックすれば参照できる (英語).

# 変数の単位の変換

元のデータの中古マンション価格 (price) は円単位.



万円単位にするには、新たな変数を作成し、元の変数を 10,000 で割ったものと定義すればよい.



## 実習 2

1. 万円単位の中古マンション価格の変数を作成する。gretl のメニューバーから「追加」→「新規変数の定義」と操作。
2. 出てきたダイアログボックスの入力ボックスに
$$\text{price\_10th}=\text{price}/10000$$
と入力し、「OK」をクリック。
  - ▶ 「price\_10th」という変数が作成され、「price を 10,000 で割ったもの」と定義される。
  - ▶ 10th は ten thousand (10,000) という意味。
3. 「id」から「price\_10th」までの7個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。

The screenshot shows a window titled "gretl: データ表示" (gretl: Data Display). The window contains a table of data. The top part of the table shows a list of years and corresponding values. The bottom part of the table shows two columns: "onekr" and "price\_10th", with values for each row from 1 to 13.

183	192	4.2e+007	0	1993	33
184	193	8e+007	12	1989	140
185	194	3.2e+007	4	1995	50
186	195	5e+007	6	2000	70
187	196	5.3e+007	4	2005	60
188	197	2.1e+007	5	2002	25
189	198	6.5e+007	8	2005	75
190	199	3.8e+007	6	2000	55
191	200	4.8e+007	6	2000	65
192	201	4.8e+007	5	2004	50
193	202	1.1e+007	4	1968	35
194	204	3.4e+007	8	1986	60
	onekr	price_10th			
1	1	620			
2	0	3700			
3	0	3700			
4	0	3400			
5	1	1400			
6	0	3000			
7	1	2900			
8	0	2900			
9	0	2800			
10	0	3200			
11	0	2900			
12	0	2900			
13	0	2700			

このような画面が表示されれば成功. onekr と price\_10th 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.

# 「築年数」変数の作成

- ▶ データセットにあるのは、「建築年」変数 (year).
- ▶ データの観測時点は 2010 年.



「築年数」変数を作成するには、新たな変数を作成し、2010 から建築年 (year) を引いたものと定義すればよい。

4. **上書き保存**する。メニューバーから「ファイル」→「データを保存」と操作すると、「データセットは現在、サブサンプルされています  
全範囲に戻しますか？」というメッセージが表示されるので、「**いいえ**」をクリック。

## 実習 3

1. 築年数の変数を作成する。gretl のメニューバーから「追加」→「新規変数の定義」と操作。
2. 出てきたダイアログボックスの入力ボックスに  
age=2010-year  
と入力し、「OK」をクリック。
  - ▶ 「age」という変数が作成され、「2010 から year を引いたもの」と定義される。
3. 「id」から「age」までの8個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。

The screenshot shows the 'gretl: データ表示' window. The top part displays a list of data points with columns for observation number, year, and values. The bottom part shows a detailed view of the first 13 rows for variables 'onekr', 'price\_10th', and 'age'.

Obs	Year	onekr	price_10th	age	
103	192	4.2e+007	0	1993	33
184	193	8e+007	12	1989	140
185	194	3.2e+007	4	1995	50
186	195	5e+007	6	2000	70
187	196	5.3e+007	4	2005	60
188	197	2.1e+007	5	2002	25
189	198	6.5e+007	8	2005	75
190	199	3.8e+007	6	2000	55
191	200	4.8e+007	6	2000	65
192	201	4.8e+007	5	2004	50
193	202	1.1e+007	4	1968	35
194	204	3.4e+007	8	1986	60

  

	onekr	price_10th	age
1	1	620	26
2	0	3700	11
3	0	3700	12
4	0	3400	9
5	1	1400	18
6	0	3000	1
7	1	2900	1
8	0	2900	1
9	0	2800	1
10	0	3200	1
11	0	2900	1
12	0	2900	1
13	0	2700	1

このような画面が表示されれば成功. onekr と price\_10th と age 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.

4. **上書き保存**する。メニューバーから「ファイル」→「データを保存」と操作すると、「データセットは現在、サブサンプルされています  
全範囲に戻しますか？」というメッセージが表示されるので、「**いいえ**」をクリック。

# 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、「**全ての統計量**を表示する」を選んで「OK」をクリックすると、選んだ変数の様々な統計量が表示される。



# 「全ての統計量を表示する」での統計量

## ▶ 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.

## ▶ 標準偏差

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

## ▶ 最小値

$$\min\{x_i\}.$$

## ▶ 最大値

$$\max\{x_i\}.$$

▶ 変動係数

▶  $CV_x = \frac{S_x}{\bar{x}}$ .

▶ 歪度

- ▶ (この授業のレベルを超えるので省略)

▶ 過剰尖度

- ▶ (この授業のレベルを超えるので省略)

▶  $\alpha$  百分位数

- ▶ 観測値を小さい順に並べたときに  $0.01\alpha n$  番目に来る値.
- ▶ 50 百分位数は中央値と同じ.
- ▶ gretl では 5 百分位数と 95 百分位数が出力できる.

- ▶ IQ 幅: 四分位範囲 (Interquartile range)
  - ▶ 75 百分位数 - 25 百分位数.
- ▶ 欠損値数
  - ▶ 値が観測されていない個体 (その変数において空白になっている個体) の数.

## 実習 4

欠損のある個体の消去，新規変数の作成を行った後のデータセットで記述統計を出力し，教科書『新しい計量経済学』 p.13 の表 1.3 の再現を試みる．

1. Ctrl キーを押しながら「minutes」，「area」，「onekr」，「price\_10th」，「age」の5つを左クリックして選択し，その上で右クリック→「基本統計量」と操作．
2. 「**全ての統計量を表示する**」を選んで状態で「OK」をクリックすると，選択した変数の記述統計 12 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．**

	平均	中央値	最小値	最大値
minutes	8.9845	8.0000	0.00000	29.000
area	53.531	50.000	10.000	280.00
onekr	0.19072	0.00000	0.00000	1.0000
price_10th	3762.6	3600.0	500.00	19000
age	15.201	11.500	1.0000	43.000

  

	標準偏差	変動係数	歪度	過剰尖度
minutes	5.4130	0.60248	0.79320	0.37659
area	29.115	0.54390	2.5473	17.388
onekr	0.39389	2.0652	1.5745	0.47891
price_10th	2151.0	0.57167	1.9703	11.370
age	11.406	0.75035	0.77453	-0.48386

  

	5百分位数	95百分位数	10幅	欠損値数
minutes	1.7500	19.000	7.0000	0
area	15.000	95.000	35.000	0
onekr	0.00000	1.0000	0.00000	0
price_10th	820.00	7000.0	2525.0	0
age	1.0000	38.000	18.250	0

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、1段目は左から平均，中央値，最小値，最大値の順。2段目は左から標準偏差，変動係数，歪度，過剰尖度の順。3段目は左から5百分位数，95百分位数，IQ幅（四分位範囲），欠損値数の順。
- ▶ 実習1で欠損値を消去したので，今のデータセットでは5変数全てについて，欠損値数は0となっている。

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. summary20201023.txt という名前で2020microdatag フォルダに保存. すると、表示された記述統計をそのままテキストファイルで保存できる.

# 教科書との数値の違い

- ▶ age (築年数) の平均は 15.201, 標準偏差は 11.406, 最小値は 1 となっていたが, 教科書では平均が 14.99, 標準偏差が 11.49, 最小値が 0 である.
  - ➡ 教科書の著者が, 1989 年建築のマンションの築年数を 21 とすべきところ, 誤って 0 としたため (付録データにて確認).

本日の作業はここまで.