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）．

$$\Downarrow$$

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.

$$\Downarrow$$

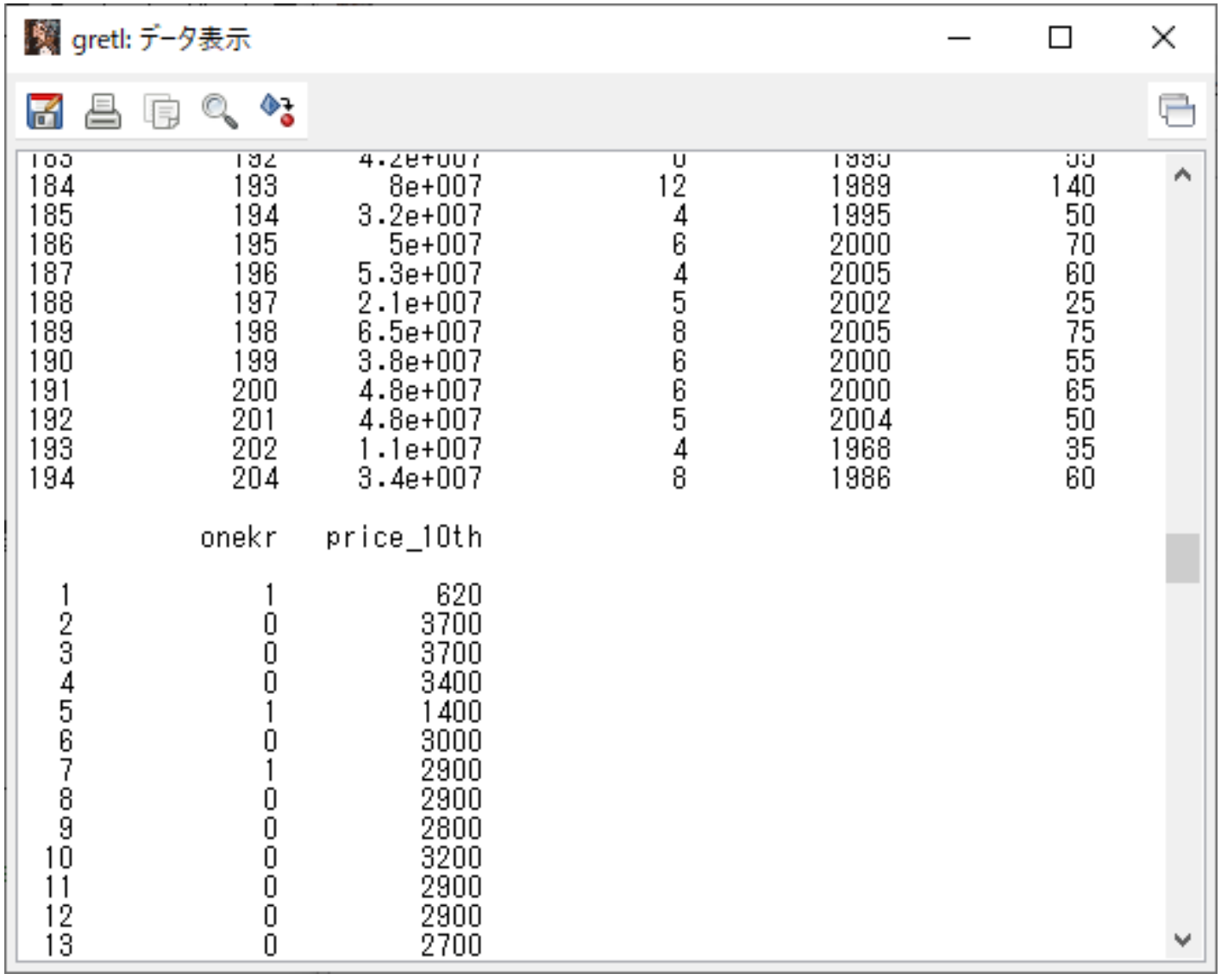
万円単位にするには，新たな変数を作成し，元の変数を 10,000 で割ったものと定義すればよい.

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

   `price_10th=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.



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.

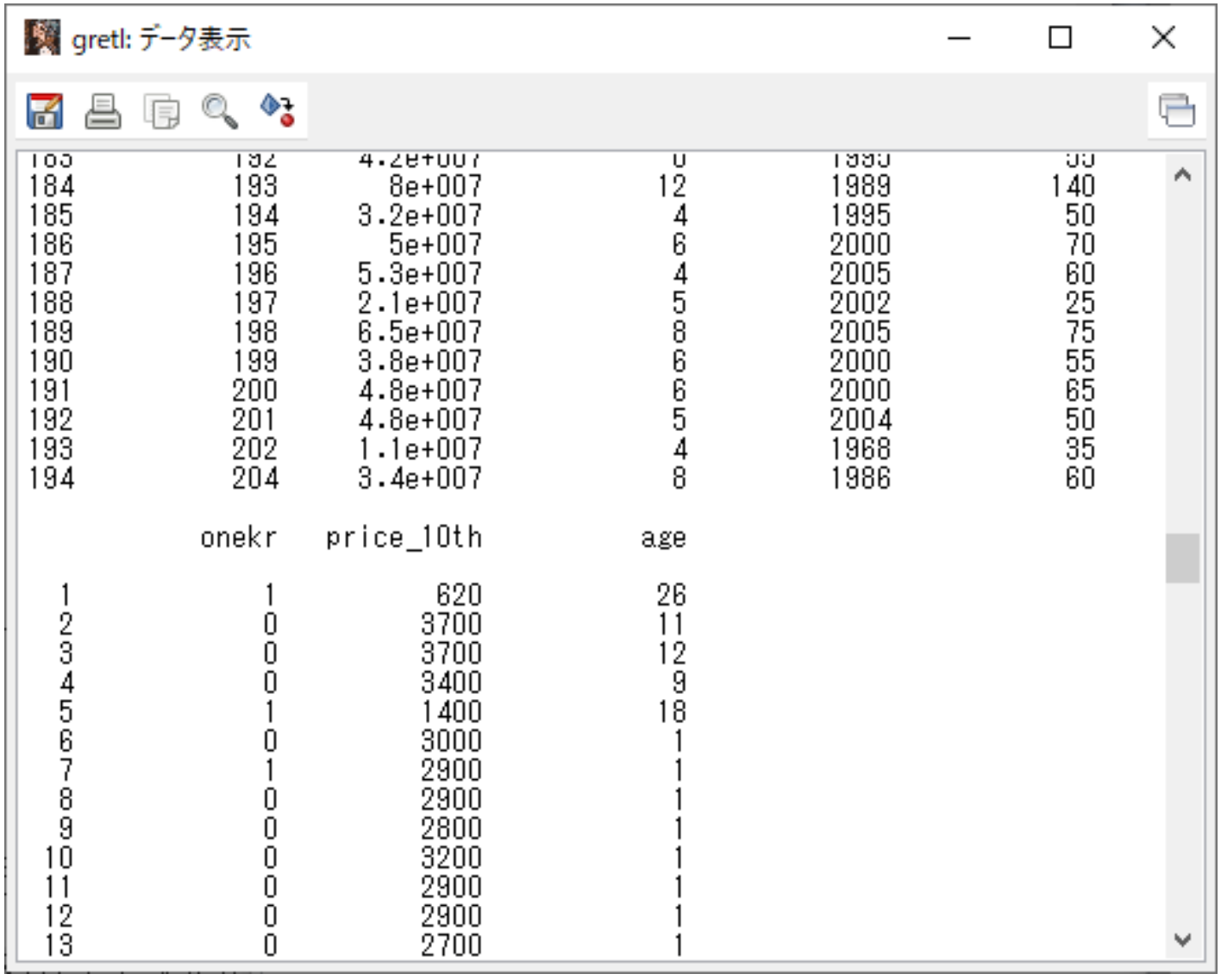
4. 上書き保存する．メニューバーから「ファイル」→「データを保存」と操作すると，「データセットは現在、サブサンプルされています 全範囲に戻しますか？」というメッセージが表示されるので，「いいえ」をクリック．

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.



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
  - $\max\{x_i\}.$

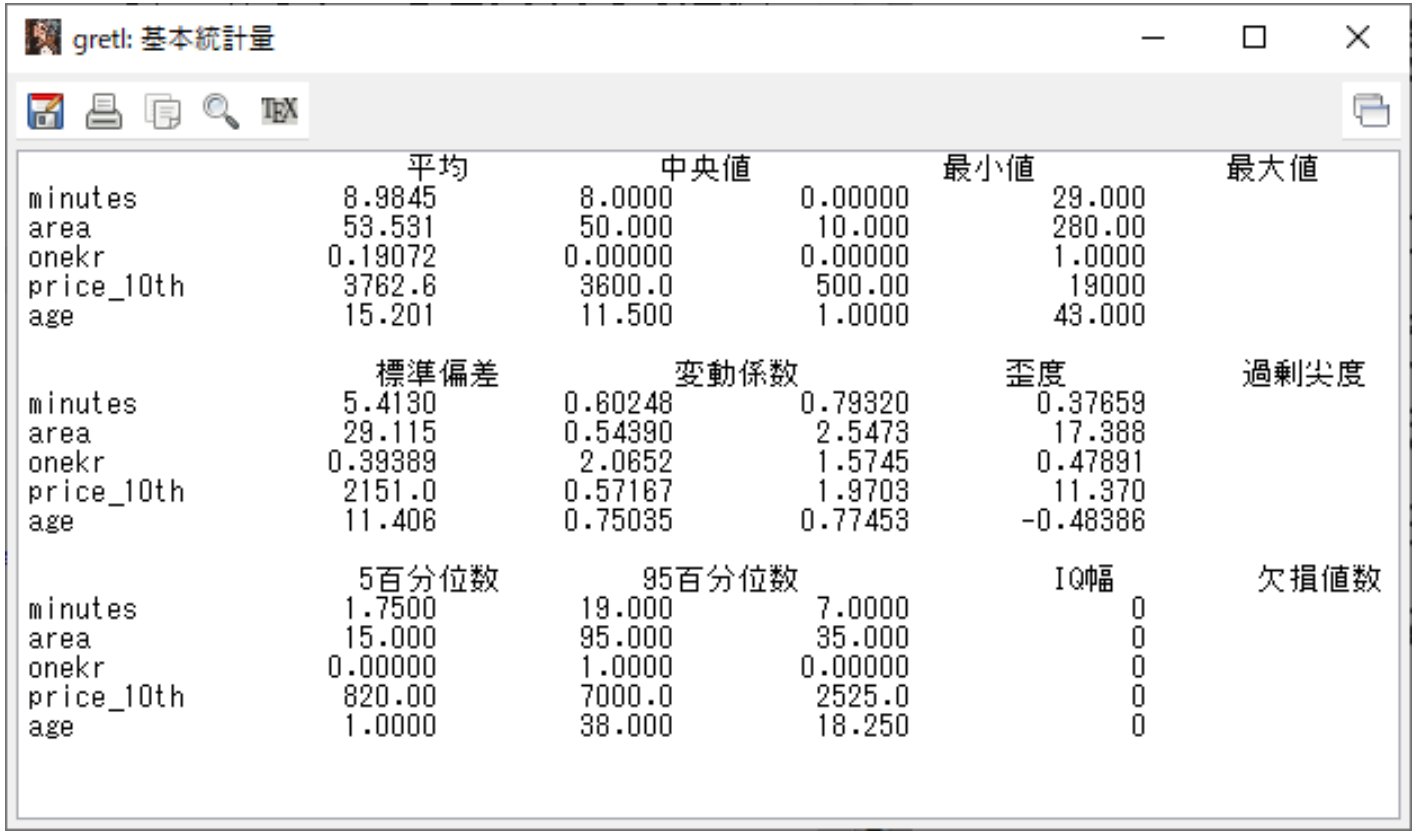
- 変動係数
  - $cv_x = \dfrac{s_x}{\bar{x}}$.
- 歪度
  - （この授業のレベルを超えるので省略）
- 過剰尖度
  - （この授業のレベルを超えるので省略）
- α 百分位数
  - 観測値を小さい順に並べたときに $0.01\alpha n$ 番目に来る値.
  - 50 百分位数は中央値と同じ.
  - gretl では 5 百分位数と 95 百分位数が出力できる.

- 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．



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.