第 11 回：ダミー変数を含む回帰

【教科書第 7 章】

北村　友宏

2020 年 12 月 18 日

本日の内容

1. ダミー変数を含む回帰

2. 重回帰モデル推定結果表の作成

ダミー変数を含む回帰

線形回帰モデル

$$
\begin{aligned}
y_i &= \beta_0 + \beta_X x_i + \beta_D d_i + u_i, \\
E(u_i \mid x_i, d_i) &= 0, \\
E(u_i u_j \mid x_i, d_i) &= 0 \quad (i \neq j), \\
V(u_i \mid x_i, d_i) &= \sigma^2, \\
i &= 1, 2, \cdots, n
\end{aligned}
$$

を推定することを考える.

- d_i はダミー変数（0 と 1 の値のみをとる）.
  - e.g., ワンルームダミー
    （ワンルーム＝ 1，それ以外＝ 0）

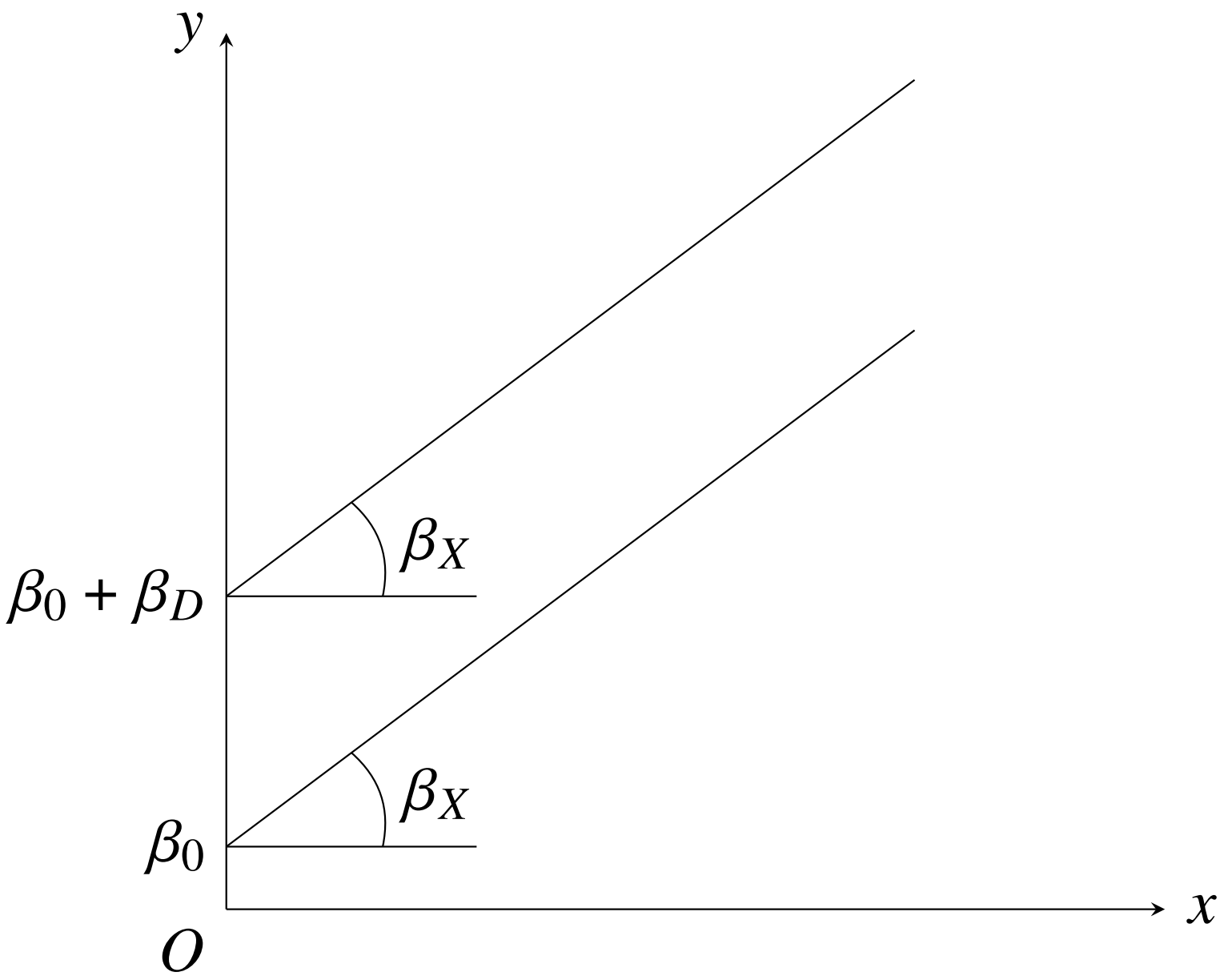
- $d_i = 0$ のとき

$$y_i = \beta_0 + \beta_X x_i + u_i = \underbrace{\beta_0}_{\text{切片}} + \beta_X x_i + u_i.$$

- $d_i = 1$ のとき

$$\begin{aligned} y_i &= \beta_0 + \beta_X x_i + \beta_D + u_i \\ &= \underbrace{(\beta_0 + \beta_D)}_{\text{切片}} + \beta_X x_i + u_i. \end{aligned}$$

$\Rightarrow$ ダミー変数の値が 0 か 1 かによって，縦軸切片が変化する．
$\Rightarrow$ ダミー変数の偏回帰係数 β_D の OLS 推定値 $\hat{\beta}_D$ を求めれば，$d_i = 1$ の場合は $d_i = 0$ の場合と比べて y_i がどの程度変化するかが分かる．


y
$\beta_0 + \beta_D$
β_X
β_0
β_X
x
O

説明変数にダミー変数を含む場合の注意

- すべての個体について，ダミー変数の値の合計が 1 になるような複数のダミー変数を作成した場合は，そのうち 1 つを除外して説明変数に用いる．
  - e.g.,「ワンルームダミー＋その他ダミー＝ 1」
    ➥ ワンルームダミーかその他ダミーどちらか 1 つを説明変数に用いる．

- 除外したダミー変数が表すものを基準として，ダミー変数の（偏）回帰係数は基準と比較してどの程度，被説明変数に対する影響度合いが異なるか，という解釈．
  - e.g., その他ダミーを除外してワンルームダミーを説明変数に用いた場合，ワンルームのマンションはそれ以外の種類のマンションと比べて被説明変数がどの程度異なるかが分かる．

ダミー変数の合計

id	onekr	other	onekr+other
1	1	0	1
2	0	1	1
4	0	1	1
5	0	1	1
6	1	0	1
7	0	1	1
8	1	0	1
9	0	1	1

$\Downarrow$

onekr というダミー変数と other というダミー変数が完全に相関する.

- 除外するダミー変数（基準）を変更しても，残る全てのダミー変数を説明変数として用いる限り，他の説明変数の偏回帰係数の推定値や標準誤差は変わらない.

- gretl では，すべての個体についてダミー変数の値の合計が 1 になるようなダミー変数を全て説明変数に選んだ場合，ダミー変数のうち 1 つが自動的に除外されて結果が表示される.

gretl でのダミー変数を含む回帰分析

$$price_i = \beta_0 + \beta_1 minutes_i + \beta_2 age_i + \beta_3 area_i + \beta_4 d_i + u_i$$

- $price_i$: 中古マンション価格（万円）
- $minutes_i$: 最寄り駅までの所要時間（分）
- age_i : 築年数（年）
- $area_i$: 面積（m^2）
- d_i : ワンルームダミー
- i : 中古マンション番号

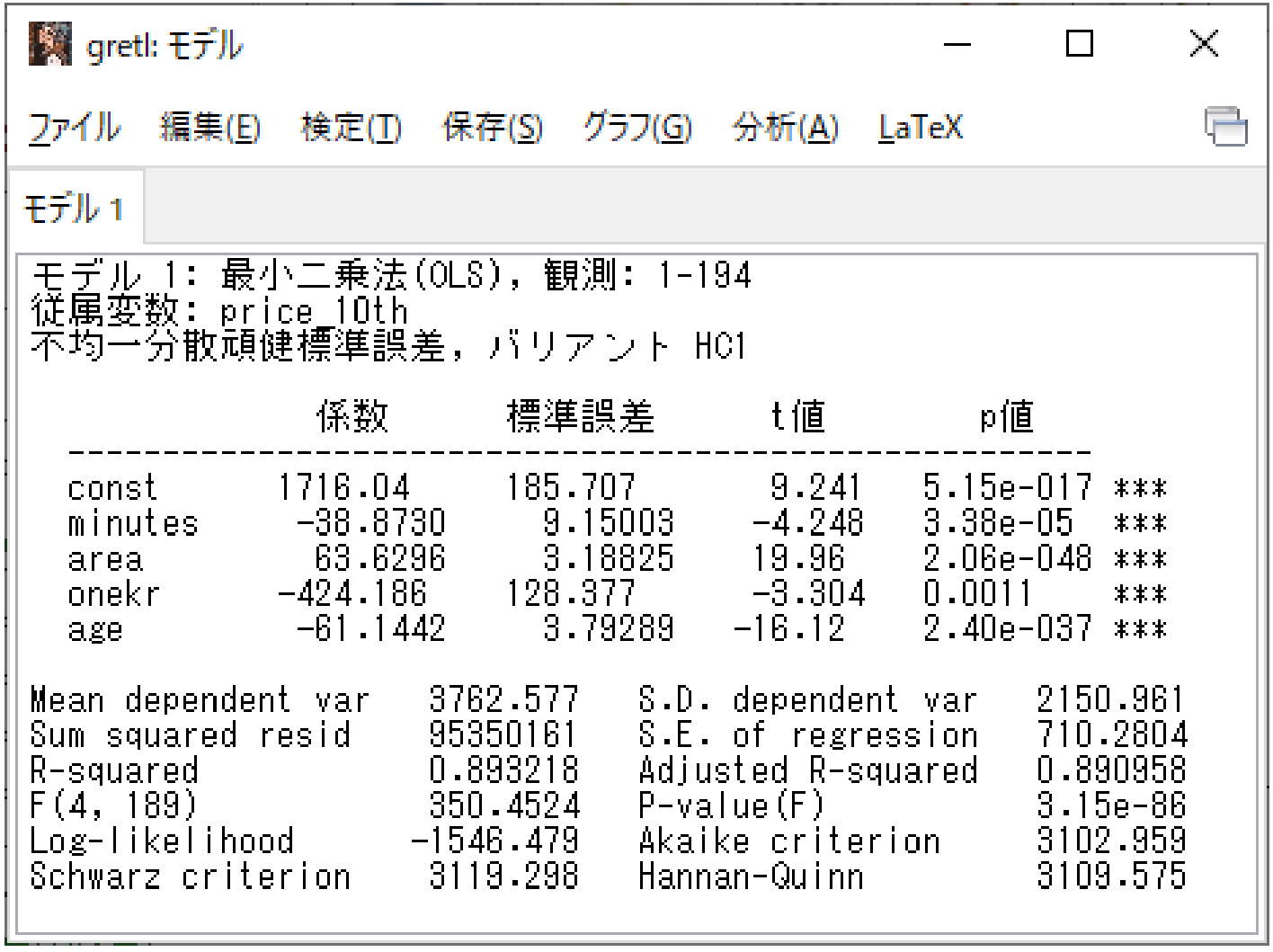
を推定する.

実習 1

まず，モデル 1（価格を所要時間のみに回帰）を推定する.

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「モデル」→「通常の最小二乗法」と操作.
5. 出てきたウィンドウ左側の変数リストにある price_10th をクリックし，3 つの矢印のうち上の青い右向き矢印をクリック.
   - 推定式の左辺の変数（被説明変数，従属変数）が price_10th（万円単位の中古マンション価格）となる.

6. 「デフォルトとして設定」にチェック.
   - gretl を終了するまでの間，次回以降「通常の最小二乗法」での推定を行う際に，いま選択した変数が自動的に被説明変数（従属変数）に入力される.
7. ウィンドウ左側の変数リストにある minutes をクリックした後，Ctrl キーを押しながら，area, onekr, age をクリックし，3 つの矢印のうち真ん中の緑の右向き矢印をクリック.
   - 推定式の右辺の変数（説明変数，独立変数）が minutes（最寄り駅までの所要時間），area（面積），onekr（ワンルームダミー），age（築年数）の 4 つとなる.
   - 最初から説明変数リストに入っている const は推定式の切片（定数項）のこと.
8. 「頑健標準誤差を使用する」にチェック．これで，推定式の誤差項 u_i のバラつき（分散）に関する仮定が誤っていても，より厳密な分析ができるようになる.



このような画面が表示されれば成功．まだ作業があるので，「gretl: モデル」のウィンドウはまだ閉じない！

モデル推定結果

- ワンルームダミーの係数
  - −424.186（符号は負）
  - 有意水準 1%で係数ゼロの H_0 棄却.
    ➥ ワンルームであるかどうかはマンションの価格と統計的に有意に相関している.
    ➥ 最寄り駅までの所要時間，築年数，面積を一定とした上で，ワンルームマンションはそれ以外の種類のマンションに比べ，市場価値が 424.186 万円安い傾向がある.

- 最寄り駅所要時間の係数
  - −38.873（符号は負）
  - 有意水準 1%で係数ゼロの H_0 棄却.
    ➥ 最寄り駅までの所要時間はマンションの価格と統計的に有意に相関している.
    ➥ 築年数，面積，ワンルームかどうかを一定とした上で，最寄り駅までの所要時間が 1 分長くなると，マンションの市場価値が 38.873 万円安くなる傾向がある.

- 自由度修正済み決定係数
  - $\bar{R}^2 = 0.890958.$
    ➥ 最寄り駅までの所要時間，築年数，面積，ワンルームかどうかの違いで，「価格」のバラつきが約 89.1%説明できている.

実習 2

1. Word を起動し，results20201218.docx という名前で 2020microdatag フォルダに保存.
2. 「挿入」→「表」と操作して 7 行 4 列の表を作る.
3. 表全体をドラッグし，「参照設定」→「図表番号の挿入」と操作.
4. ラベルを「表」に，位置を「選択した項目の上」して OK をクリックすると，表のすぐ上の行に「表 1」と入力される.
   - ラベルに「表」がなければ，「新しいラベル...」をクリックして出てくるダイアログボックスの入力ボックスに表と入力して OK をクリック.
5. 「表 1」の後に全角スペースを入れてモデル推定結果と入力し，中央揃えにする.

6. 表の１行２列目に偏回帰係数，１行３列目に標準誤差と入力.
7. 表の１行２列目から１行３列目までをドラッグし，「レイアウト」タブ（右端の，色が濃いほう）から「配置」→「下揃え（中央）」と操作.
8. 表の１列目に，以下のように入力.
   - ２行１列目：最寄り駅所要時間
   - ３行１列目：築年数
   - ４行１列目：面積
   - ５行１列目：ワンルームダミー
   - ６行１列目：定数項
   - ７行１列目：自由度修正済み決定係数
9. 表の２行１列目から７行１列目までをドラッグし，「レイアウト」タブ（右端の，色が濃いほう）から「配置」→「中央揃え」と操作.

10. gretl に出力されていた推定結果の数値を，Word で作成した表の対応するセルにコピー・貼り付けする．数値をドラッグして選択し，メニューバーから「編集」→「コピー」と操作すればコピーできる．
    - const は定数項，minutes は最寄り駅所要時間，age は築年数，area は面積，onekr はワンルームダミー，係数は偏回帰係数，Adjusted R-squared は自由度修正済み決定係数．
    - 自由度修正済み決定係数は，偏回帰係数の列（7 行 2 列目）に入力する．
11. 偏回帰係数，標準誤差，自由度修正済み決定係数は小数第 3 位を四捨五入．

12. 頑健標準誤差を用いた結果を見て，不均一分散に対して頑健な標準誤差の右隣のセルには，その変数の係数の p 値が 0.01 未満なら***，0.05 未満なら**，0.10 未満なら*と入力する.
13. 表の 2 行 2 列目から 7 行 3 列目までをドラッグし，「レイアウト」タブ（右端の，色が濃いほう）から「配置」→「下揃え（右）」と操作.
14. 表の 2 行 4 列目から 6 行 4 列目までをドラッグし，「レイアウト」タブ（右端の，色が濃いほう）から「配置」→「下揃え（左）」と操作.

15. Word で作成した表のすぐ下の行に，

（注 1）表中の***は有意水準 1%で統計的に有意であることを表す.
（注 2）不均一分散に対して頑健な標準誤差を用いている.
（注 3）観測値数は 194 である.

と入力して上書き保存.

- 「アスタリスク 1 つ（有意水準 10%）と 2 つ（有意水準 5%)」はこの表では出てこなかったので省略.
- 観測値数は，出力結果の「観測数」と記載されている箇所を見れば分かる.

教科書との数値の違い

教科書の中古マンションデータと同じデータを使って分析をしたはずだが，モデル推定結果が，今回 gretl で出力したものと教科書 p.130 の推定結果で異なっている．

- e.g., ワンルームダミーの係数推定値が，小数第 3 位を四捨五入すると −424.19 となっていたが，教科書では −544.81 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）．

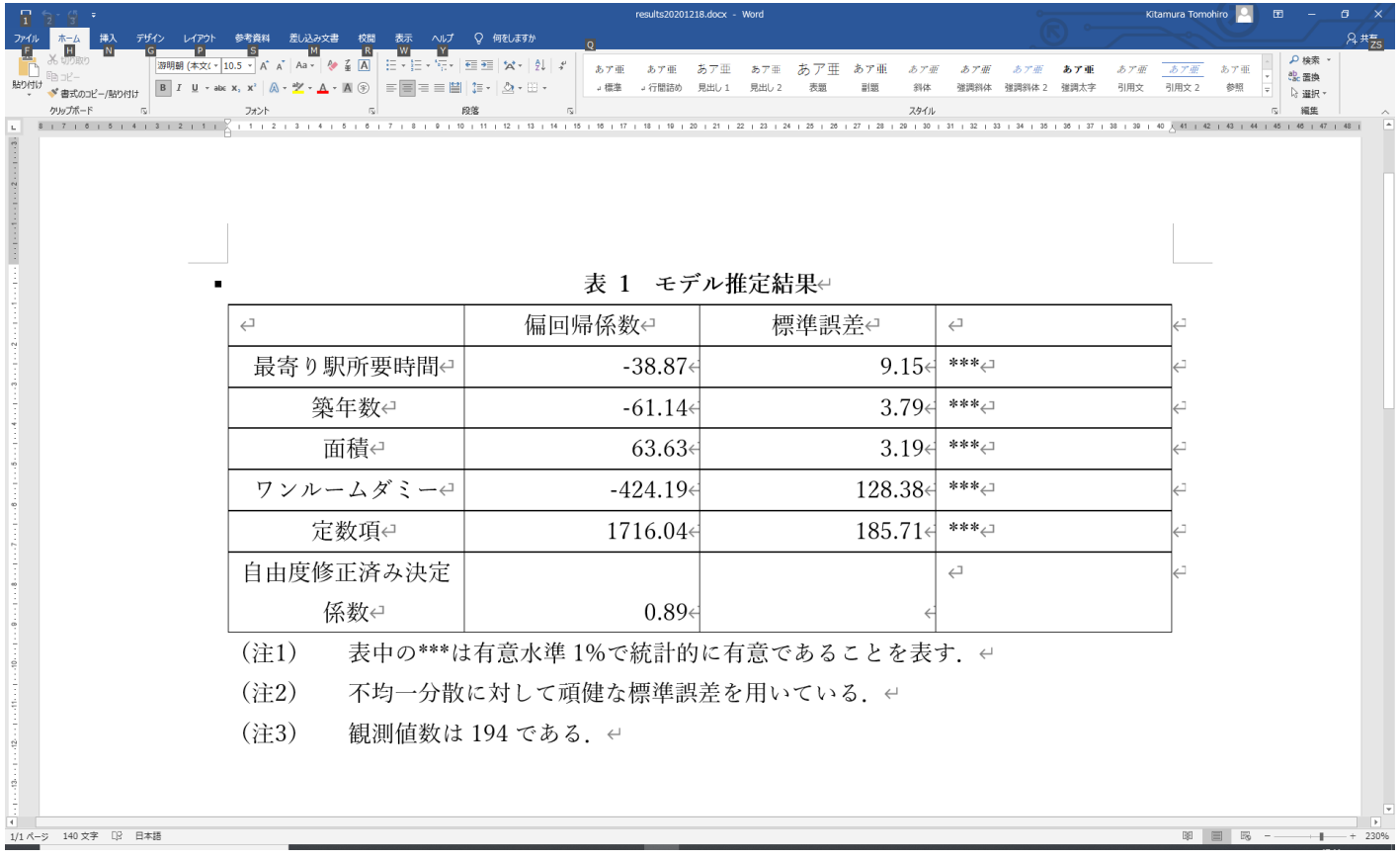


表 1　モデル推定結果

	偏回帰係数	標準誤差	
最寄り駅所要時間	-38.87	9.15	***
築年数	-61.14	3.79	***
面積	63.63	3.19	***
ワンルームダミー	-424.19	128.38	***
定数項	1716.04	185.71	***
自由度修正済み決定係数	0.89		

（注1）　表中の***は有意水準 1%で統計的に有意であることを表す．

（注2）　不均一分散に対して頑健な標準誤差を用いている．

（注3）　観測値数は 194 である．

このような表を作成できればよい．

本日の作業はここまで.
今回は gretl のデータセットに変更を加えていないので，gretl のデータセット（setagayaapartment.gdt）を上書き保存する必要はない.