第 9 回：重回帰モデルの係数の検定

【教科書第 6 章】

北村　友宏

2020 年 12 月 4 日

本日の内容

1. 重回帰モデル
2. 欠落変数バイアス
3. gretl での重回帰分析

重回帰

- 定数項以外に説明変数が複数ある回帰モデルを重回帰モデル（multiple regression model）という．

重回帰モデルを，ベクトル・行列を用いて簡潔に表すと，

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \\ E(\boldsymbol{u} \mid \boldsymbol{X}) &= \boldsymbol{0}, \\ V(\boldsymbol{u} \mid \boldsymbol{X}) &= \sigma^2 \boldsymbol{I}_n. \end{aligned}$$

OLS 推定における仮定（重回帰の場合）

- 説明変数を所与として，誤差項の期待値はゼロ.
  - $E(\boldsymbol{u} \mid \boldsymbol{X}) = \boldsymbol{0}$.

  $\Longrightarrow$ 説明変数と誤差項は無相関.

- 説明変数を所与として，誤差項の分散は一定で，異なる個体の誤差項同士は無相関.
  - $V(\boldsymbol{u} \mid \boldsymbol{X}) = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \boldsymbol{I}_n.$

- 説明変数を所与として，誤差項は正規分布に従う.
  - $\boldsymbol{u} \mid \boldsymbol{X} \sim N\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n\right)$.

偏回帰係数

- 重回帰モデルの回帰係数を偏回帰係数（partial regression coefficient）という.
- 重回帰モデル

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i,$$

の偏回帰係数 β_j $(j = 1, 2, \cdots, k)$ は，「仮に x_{ij} 以外の変数を一定水準に固定したときに，$x_{i1}, x_{i2}, \cdots, x_{ik}$ を所与とした y_i の期待値に x_{ij} が与える影響」を測る.

  - e.g., 仮に中古マンションの面積が全マンションで一定だったときの，最寄り駅までの所要時間がマンションの価格の条件付き期待値に与える影響.
  - 経済学では「他の条件を一定として（*ceteris paribus*）」と表現.

- y_i の条件付き期待値をとった

$$E(y_i \mid x_{i1}, x_{i2}, \cdots, x_{ik}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik},$$

を x_{ij} で偏微分（他の説明変数の値は一定）すると，β_j になる．

$$\frac{\partial E(y_i \mid x_{i1}, x_{i2}, \cdots, x_{ik})}{\partial x_{ij}} = \beta_j.$$

- x_{ij} が y_i に与える影響に興味がある場合，「その他の変数の影響を一定」という状況を作り出すための，x_{ij} 以外の説明変数はコントロール変数．

欠落変数バイアス

真のモデルは

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + w_i,$$
$$E(w_i \mid x_{i1}, x_{i2}) = 0,$$

であるが，x_{i1} が y_i に与える影響に興味があるために，x_{i2} を除外して

$$y_i = \beta_0 + \beta_1 x_{i1} + u_i,$$

を推定する，つまり y_i を x_{i1} のみに単回帰することを考える．

- $u_i = \beta_2 x_{i2} + w_i$.

もし x_{i1} と x_{i2} が相関していると，$u_i = \beta_2 x_{i2} + w_i$ なので x_{i1} と u_i も相関する．つまり，

$$\mathrm{Cov}(x_{i1}, u_i) \neq 0.$$

$$\Downarrow$$

y_i を x_{i1} のみに回帰すると，x_{i1} の係数の OLS 推定量 $\hat{\beta}_1$ は，

$$\operatorname*{plim}_{n\to\infty} \hat{\beta}_1 = \beta_1 + \frac{\mathrm{Cov}(x_{i1}, u_i)}{V(x_{i1})} \neq \beta_1.$$

$$\Downarrow$$

偏り（バイアス）が生じ，正しく推定できない（一致推定量が得られない）．

- このバイアスは，y_i に影響を与える x_{2i} が説明変数から欠落していることにより生じる．
- 説明変数の欠落によって生じる OLS 推定量の偏りを欠落変数バイアス（omitted variable bias）という．
  - 除外変数バイアスともいう．
- 欠落変数バイアスは，モデルに必要な説明変数が 1 つでも欠落している限り必ず発生する．
  $\Longrightarrow$ 通常，避けられない．
- 欠落変数バイアスを緩和する方法
  - 重回帰分析をし，他の変数の影響をコントロールする．
  - パネルデータを利用し，固定効果モデルを仮定する（この授業で用いている中古マンションのデータでは不可能）．

何をコントロールすべきか？

先行研究を参考にすればよい.

$\Uparrow$

- 各分野では，使うべきコントロール変数が定着している.
  - 物件価格の分析：物件面積
  - 労働者賃金の分析：性別，年齢，学歴
  - 子どもの学力の分析：親の年収

gretl での重回帰分析

「駅へのアクセスのよさがマンション価値に与える影響」を分析するためのモデル

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

- $price_i$: 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）
- i : 中古マンション番号

を推定する.

教科書第 6 章 4 節と同様に，以下の 3 通りのモデルを推定する．

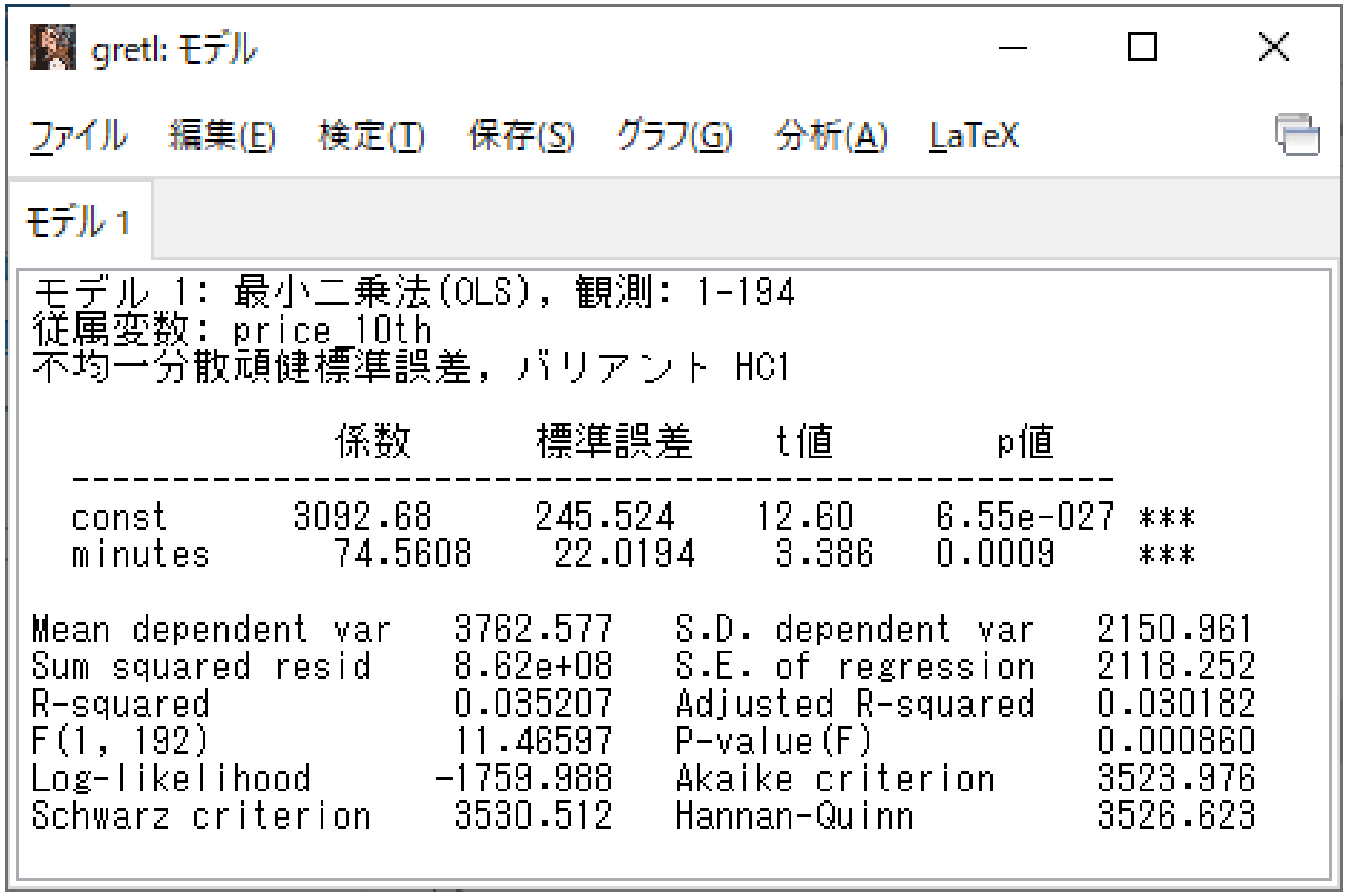
- モデル 1：価格を所要時間のみに回帰（単回帰）．
  - 「所要時間」以外の要因はコントロールしない．
  - $\beta_2 = \beta_3 = 0$ と仮定．
- モデル 2：価格を所要時間と築年数に回帰．
  - 「築年数」をコントロール．
  - $\beta_3 = 0$ と仮定．
- モデル 3：価格を所要時間，築年数，面積に回帰．
  - 「築年数」と「面積」をコントロール．

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，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）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.
9. 「OK」をクリックすると，結果が新しいウィンドウに表示される.



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！

10. 出力された「gretl: モデル」のウィンドウのメニューバーから「ファイル」→「名前を付けて保存」と操作.
11. 「標準テキスト」を選び,「OK」をクリック.
12. results20201204_1.txt という名前で 2020microdatag フォルダに保存. すると，表示された推定結果をそのままテキストファイルで保存できる.

出力結果の見方

- 係数: （偏）回帰係数推定値
- 標準誤差: （偏）回帰係数の標準誤差
- t 値: 「（偏）回帰係数が 0」という帰無仮説の両側 t 検定における検定統計量の実現値（t 値）
- p 値: 両側 p 値
- R-squared: 決定係数
- Adjusted R-squared: 自由度修正済み決定係数

自由度修正済み決定係数

- 決定係数 R^2 は説明変数の数（推定するパラメータの数）を増やすと必ず上昇する.
  ➥ 関係のない説明変数を追加しても R^2 は上昇する.
  ➥ それを回避するには，R^2 を修正する.

自由度修正済み決定係数（adjusted R-squared）は，

$$\bar{R}^2 = 1 - \left(1 - R^2\right) \cdot \frac{n-1}{n-k-1}.$$

- $\bar{R}^2$ はマイナスになることがある.
- 「重回帰の場合」や「単回帰と重回帰の結果を比較する場合」は，自由度修正済み決定係数 $\bar{R}^2$ を見るのが一般的.

標準誤差（ベクトル・行列表示）

- 推定量の標準偏差の推定値を標準誤差（standard error）という．
- j 番目の（偏）回帰係数の OLS 推定量 $\hat{\beta}_j$ の（デフォルトの）標準誤差は，

$$\text{s.e.}\left(\hat{\beta}_j\right) = \sqrt{\left[\frac{\boldsymbol{e}'\boldsymbol{e}}{n-k-1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{j,j}}.$$

$\Rightarrow$ この標準誤差は，任意の i について $V(u_i \mid \boldsymbol{X})$ が一定（均一分散）の場合のみ正しい．

頑健標準誤差

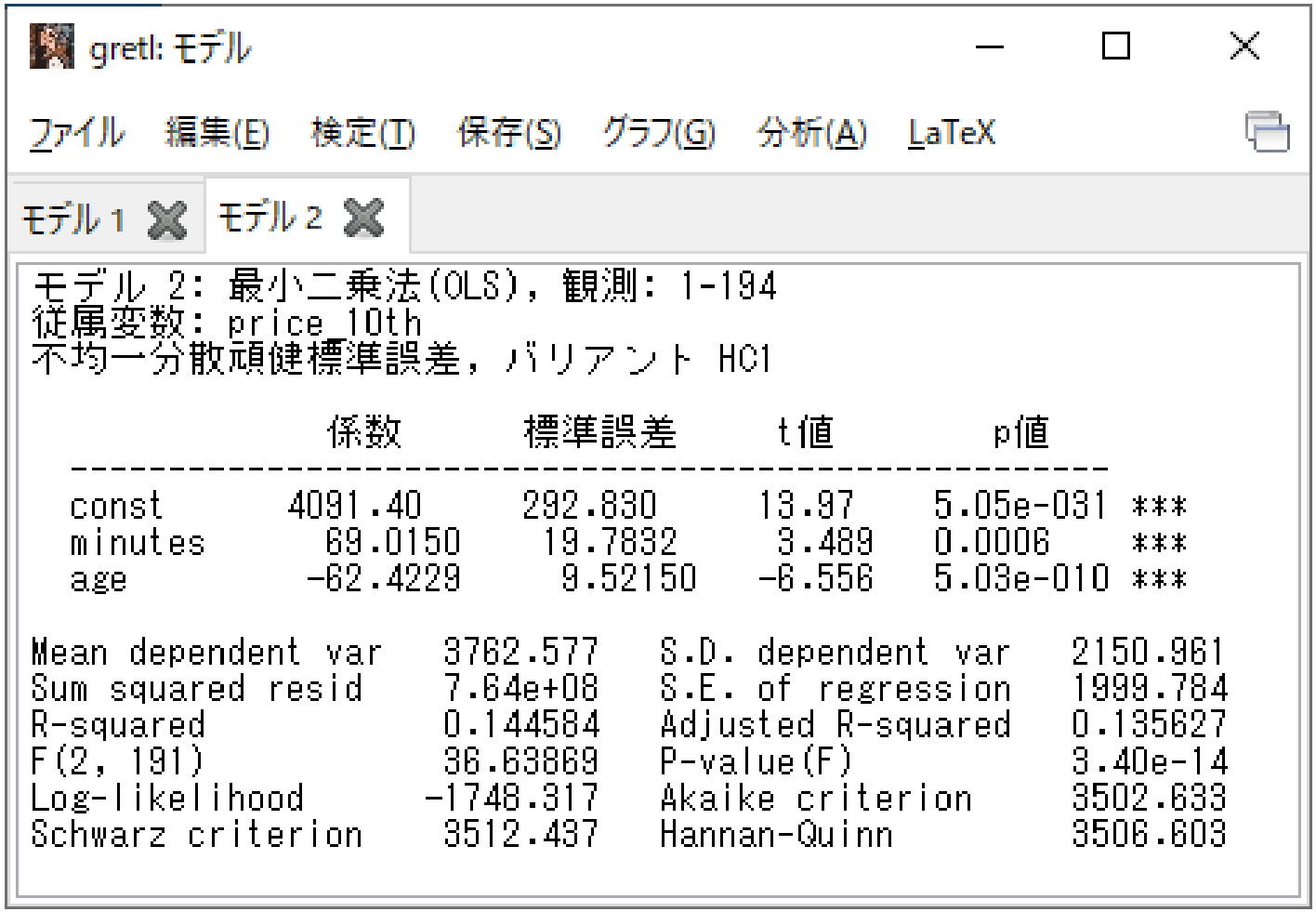
- $V(u_i \mid \boldsymbol{X})$ が一定でないことを（条件付き）不均一分散（heteroskedasticity）という．
- 不均一分散があっても厳密な標準誤差を求めるために，頑健標準誤差（robust standard error）が開発されている．

実習 2

続いて，モデル 2（価格を所要時間と築年数に回帰）を推定する．

1. gretl のメニューバーから「モデル」→「通常の最小二乗法」と操作．説明変数（独立変数）は必ず前回の選択内容が記録されており，被説明変数（従属変数）は前回「デフォルトとして設定」にチェックしていれば前回の選択内容が記録されている．
2. 従属変数の入力ボックスに price_10th が入力されていなければ，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．

3. ウィンドウ左側の変数リストにある 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）と age（築年数）の 2 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.
4. 「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.
5. 「OK」をクリックすると，結果が表示される.



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！

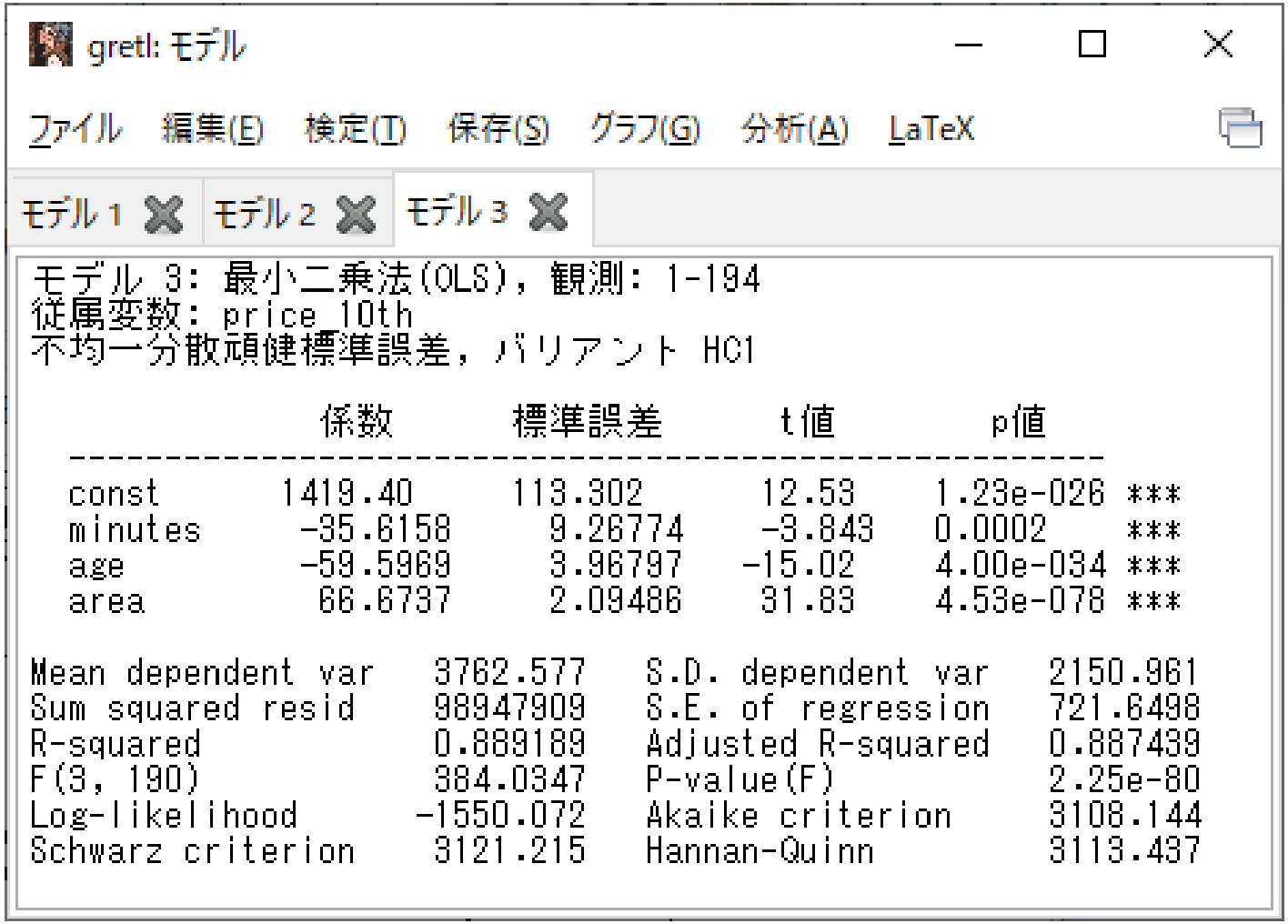
6. 「モデル 2」が表示されている状態で，「gretl: モデル」のウィンドウのメニューバーから「ファイル」→「名前を付けて保存」と操作.
7. 「標準テキスト」を選び，「OK」をクリック.
8. results20201204_2.txt という名前で 2020microdatag フォルダに保存．すると，表示された推定結果をそのままテキストファイルで保存できる.

実習 3

続いて，モデル 3（価格を所要時間，築年数，面積に回帰）を推定する．

1. gretl のメニューバーから「モデル」→「通常の最小二乗法」と操作．説明変数（独立変数）は必ず前回の選択内容が記録されており，被説明変数（従属変数）は前回「デフォルトとして設定」にチェックしていれば前回の選択内容が記録されている．
2. 従属変数の入力ボックスに price_10th が入力されていなければ，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．

3. ウィンドウ左側の変数リストにある area 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）と age（築年数）と area（面積）の 3 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．
4. 「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．
5. 「OK」をクリックすると，結果が表示される．



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！

6. 「モデル 3」が表示されている状態で，「gretl: モデル」のウィンドウのメニューバーから「ファイル」→「名前を付けて保存」と操作.
7. 「標準テキスト」を選び，「OK」をクリック.
8. results20201204_3.txt という名前で 2020microdatag フォルダに保存. すると，表示された推定結果をそのままテキストファイルで保存できる.

モデル 1 vs モデル 2

- 最寄り駅所要時間の係数
  - 74.5608 ➜ 69.0150（符号は依然として正）
  - モデル 1 でもモデルで 2 も，有意水準 1%で，係数ゼロの H_0 棄却.
    ➥ モデル 1 でもモデル 2 でも，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.

- 自由度修正済み決定係数
  - $\bar{R}^2 = 0.030182$ ➜ $\bar{R}^2 = 0.135627$.
    ➥ 当てはまりが改善され，「最寄り駅までの所要時間」と「築年数」の違いで，「価格」のバラつきが約 13.6%説明できるようになった.

モデル 2 vs モデル 3

- 最寄り駅所要時間の係数
  - 69.0150 ➜ −35.6158（符号が負になった！ ）
  - モデル 2 でもモデル 3 でも，有意水準 1%で，係数ゼロの H_0 棄却.
    ➥ モデル 2 でもモデル 3 でも，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.
    ➥ 築年数と面積を一定とした上で，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 35.6158 万円（356,158 円）安くなる傾向がある.
- 自由度修正済み決定係数
  - $\bar{R}^2 = 0.135627$ ➜ $\bar{R}^2 = 0.887439$.
    ➥ 当てはまりが大きく改善され，「最寄り駅までの所要時間」と「築年数」と「面積」の違いで，「価格」のバラつきが約 88.7%説明できるように.

面積をコントロールしたことによる符号逆転の理由

- 駅から遠い場所ほど面積の広い物件が多い.
- 面積の広い物件ほど価格が高い.

$\Rightarrow$「面積」をコントロールしなければ，部屋の広さによる価格上昇効果が拾われる.

$\Downarrow$

面積をコントロールしていない（面積が説明変数に入っていない）モデル１とモデル２では，最寄り駅所要時間の係数が正となった.

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 2・モデル 3 の推定結果が，今回 gretl で出力したものと教科書 p.112 の表 6.1 の推定結果表で異なっている．

- e.g., モデル 3 の最寄り駅所要時間の係数推定値の，小数第 3 位を四捨五入すると −35.62 となっていたが，教科書では −32.68 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.
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.