第 6 回：単回帰モデルの推定
【教科書第 4 章・第 6 章】

北村　友宏

2020 年 11 月 6 日

本日の内容

1. 単回帰モデル

2. gretl での単回帰分析

単回帰

大きさ n の２変量データ
$((y_1, x_1), (y_2, x_2), \cdots, (y_n, x_n))$ を用いて，線形回帰モデル（linear regression model）

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_i + u_i, \\
E(u_i \mid x_i) &= 0, \\
E(u_i u_j \mid x_i) &= 0 \quad (i \neq j), \\
V(u_i \mid x_i) &= \sigma^2, \\
i &= 1, 2, \cdots, n
\end{aligned}
$$

を推定することを考える．
これを推定すれば，２つの変数間の関係（x_i が増加すると y_i はどの程度変化する傾向があるか？）を定量的に検証できる．

- y_i：被説明変数（explained variable）
  - e.g., 中古マンションの価格
  - 従属変数（dependent variable）ともいう.
- x_i：説明変数（explanatory variable）
  - e.g., 中古マンションから最寄り駅までの所要時間
  - 独立変数（independent variable）ともいう.
- β_0, β_1：回帰係数（regression coefficient）
  - 特に，β_0 は定数項（constant term）.
- u_i：誤差項（error term）
  - 撹乱項（disturbance term）ともいう.

説明変数 x_i は確率的（stochastic）とする.

- 定数項以外の説明変数が 1 つである回帰モデルを単回帰モデル（simple regression model）という．

$E(u_i \mid x_i) = 0$ の仮定より，

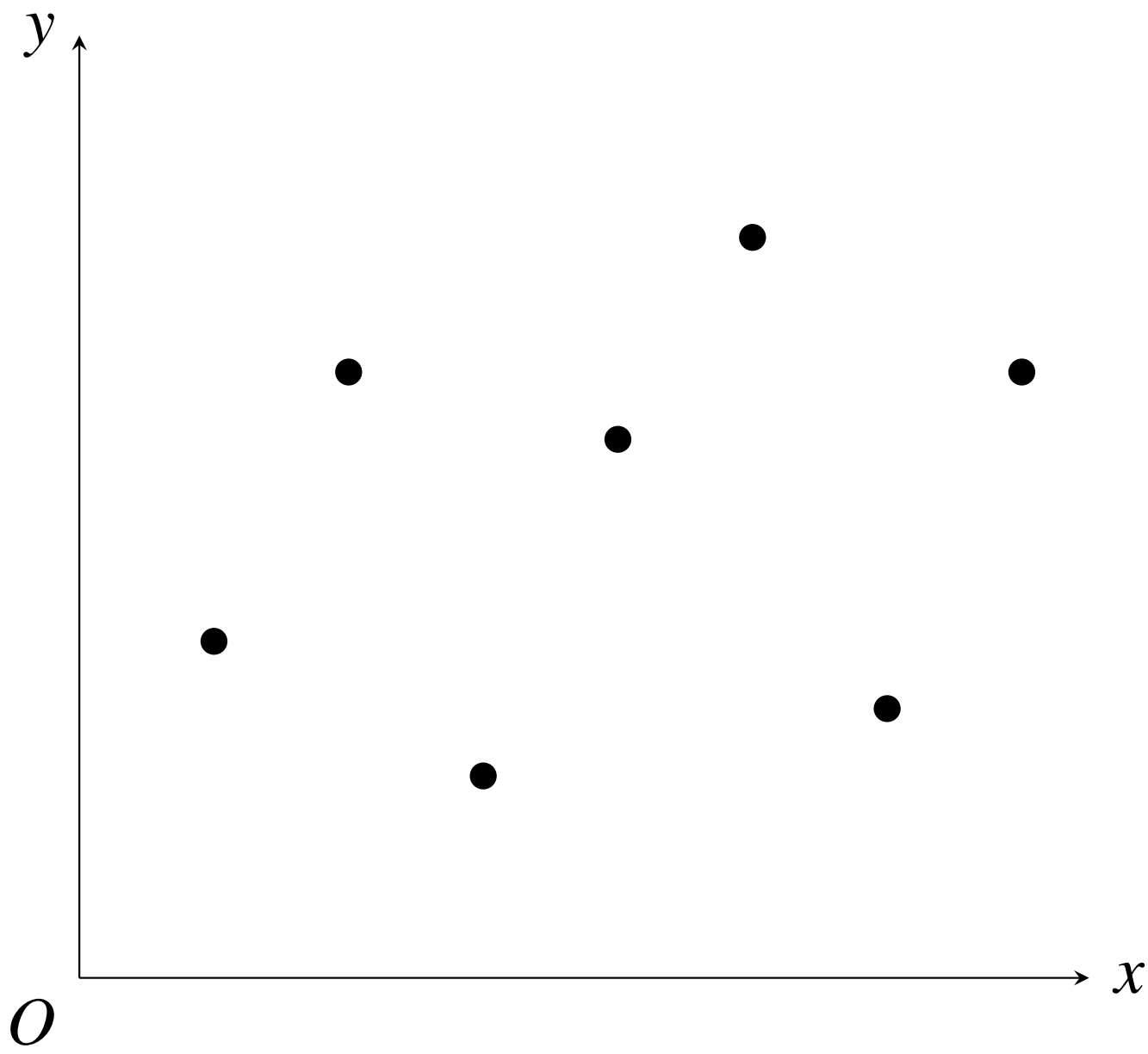
$$E(y_i \mid x_i) = \beta_0 + \beta_1 x_i.$$

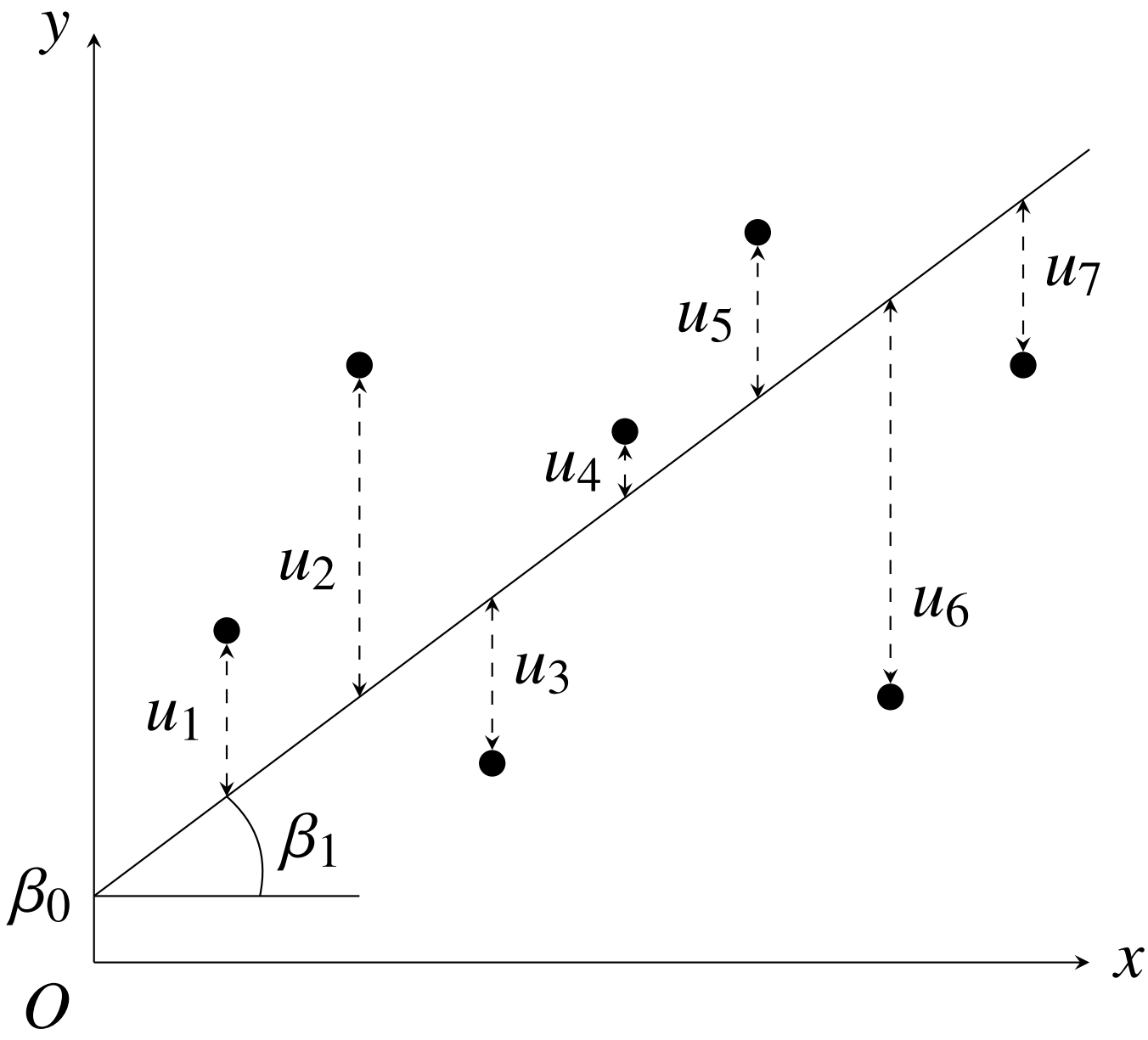
$\Rightarrow$ これは x_i が与えられたときの y_i の条件付き期待値（conditional mean）．

- $E(y_i \mid x_i)$ を求めることを，y_i を x_i に回帰する（regress）という．

$$\Downarrow$$

β_0 と β_1 を求める（推定する）には？


y
x
O


y
x
O
β_0
β_1
u_1
u_2
u_3
u_4
u_5
u_6
u_7

モデルを

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$$

と書き換え，

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

が最小になるような $\hat{\beta}_0$ と $\hat{\beta}_1$ を求める．

- e_i：残差（residual）
  - 誤差項 u_i とは別物．
- 残差二乗和 $\sum_i^n e_i^2$ が最小になるように回帰係数を求める方法を通常の最小二乗法（Ordinary Least Squares, OLS）という．

- OLS によって推定される統計量を OLS 推定量（OLS estimator）といい，その実現値を OLS 推定値（OLS estimate）という．

この場合の OLS 推定量は，

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

- $\bar{y} = \dfrac{1}{n} \sum_{i=1}^{n} y_i.$
- $\bar{x} = \dfrac{1}{n} \sum_{i=1}^{n} x_i.$

「(定数項を含む) 単回帰モデル」の OLS 推定量の導出

残差二乗和最小化問題は,

$$\min_{(\hat{\beta}_0,\hat{\beta}_1)\in\mathbb{R}^2} \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

1 階条件は,

$$\frac{\partial \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\partial \hat{\beta}_0} = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \cdot (-1) = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0, \tag{1}$$

$$\frac{\partial \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\partial \hat{\beta}_1} = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \cdot (-x_i) = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0. \tag{2}$$

(1) より，

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \Leftrightarrow \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \sum_{i=1}^{n} \hat{\beta}_1 x_i = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i = n\hat{\beta}_0$$

$$\Leftrightarrow \hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_1 \cdot \frac{1}{n}\sum_{i=1}^{n} x_i .$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

とすると，$\hat{\beta}_0$ は，

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \qquad (1')$$

(2) と (1) より，

$$\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) - \bar{x} \underbrace{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)}_{\text{(1) より，0 となる}} = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) - \sum_{i=1}^{n} \bar{x}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} \left\{ x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) - \bar{x}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \right\} = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0.$$

(1') を代入すると，

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\Leftrightarrow \sum_{i=1}^{n}(x_i - \bar{x})\left\{y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i\right\} = 0$$

$$\Leftrightarrow \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i) = 0$$

$$\Leftrightarrow \sum_{i=1}^{n}(x_i - \bar{x})\left\{y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x})\right\} = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} \left\{ (x_i - \bar{x})(y_i - \bar{y}) - (x_i - \bar{x})\hat{\beta}_1 (x_i - \bar{x}) \right\} = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) - \sum_{i=1}^{n} (x_i - \bar{x})\hat{\beta}_1 (x_i - \bar{x}) = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})^2 = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\Leftrightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

OLS 推定における仮定（単回帰の場合）

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

  $\Longrightarrow$ 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.
  - $V(u_i \mid x_i) = \sigma^2.$
  - $E(u_i u_j \mid x_i) = 0 \quad (i \neq j).$

- 説明変数を所与として，誤差項は正規分布に従う.
  - $u_i \mid x_i \sim N\left(0, \sigma^2\right).$

gretl での単回帰分析

いま整理・加工・分析している中古マンションのデータセットを用いて，
「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

$$price_i = \beta_0 + \beta_1 minutes_i + u_i$$

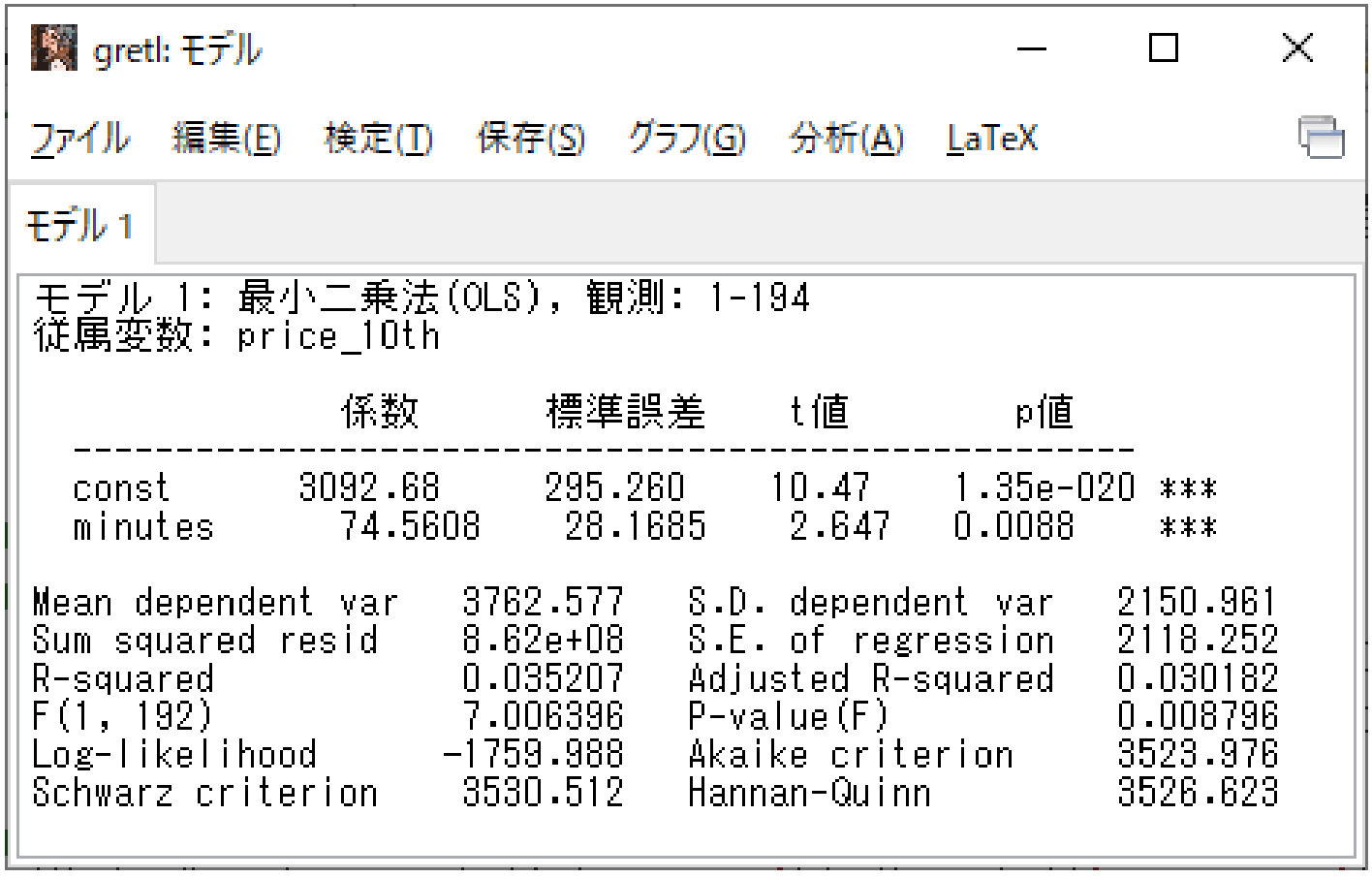
- $price_i$：中古マンション価格（万円）
- $minutes_i$：最寄り駅までの所要時間（分）
- i：中古マンション番号

を推定する．
➥「中古マンション価格」を「最寄り駅までの所要時間」に回帰する．

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「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. 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.
7. 「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！

出力結果の見方

- 係数: 回帰係数推定値
- 標準誤差: 回帰係数の標準誤差
  - 次回の授業で説明
- t 値: 「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）
  - 次回の授業で説明
- p 値: 両側 p 値
  - 次回の授業で説明
- R-squared: 決定係数

決定係数

決定係数（R-squared）は，

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}.$$

- 定数項ありの単回帰の場合，$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.
- 意味 モデルの当てはまりの良さ（説明変数で，被説明変数のバラつきのうち，どの程度の割合を説明できているか）
- $0 \leq R^2 \leq 1$.
  - $R^2 = 0$: 全く説明できていない.
  - $R^2 = 1$: 完全に説明できている.

$\Rightarrow R^2 = 0$ や $R^2 = 1$ になることは，実際の実証分析ではまず起こり得ない.

モデル推定結果

- 最寄り駅所要時間の係数
  - 74.5608（符号は正）
    ➥ 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 74.5608 万円（745,608 円）高くなる（？）
    ➥ 直感に反する.

- 定数項
  - 3092.68

- 決定係数
  - $R^2 = 0.035207.$
    ➥ 「最寄り駅までの所要時間」の違いで，「価格」のバラつきが約 3.5%のみ説明できる.

実習 2

1. 表示された「gretl: モデル 1」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。
2. 「標準テキスト」を選び，「OK」をクリック。
3. results20201106.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．

直感に反する分析結果

- 直感 駅に近いマンションほど価格が高く，駅から遠いマンションほど価格が安い.

- 分析結果 駅に近いマンションほど価格が安く，駅から遠いマンションほど価格が高い.

- 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.

$\Rightarrow$ 単に「価格」を「最寄り駅までの所要時間」に回帰すると，部屋の広さによる価格上昇効果が拾われる.

$\Downarrow$

正の係数が検出された.

- より厳密に「駅へのアクセスのよさがマンション価値に与える影響」を分析するには，マンションの面積などをコントロールする必要がある（後の授業で説明）.

本日の作業はここまで.