第３回：「ミクロデータ分析Ｉ」の復習（3）

北村　友宏

2020 年 10 月 16 日

本日の内容

1. 単回帰モデル

2. 需要関数の推定（単回帰）

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., みかんの取引数量
  - 従属変数（dependent variable）ともいう.
- x_i：説明変数（explanatory variable）
  - e.g., みかんの価格
  - 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，

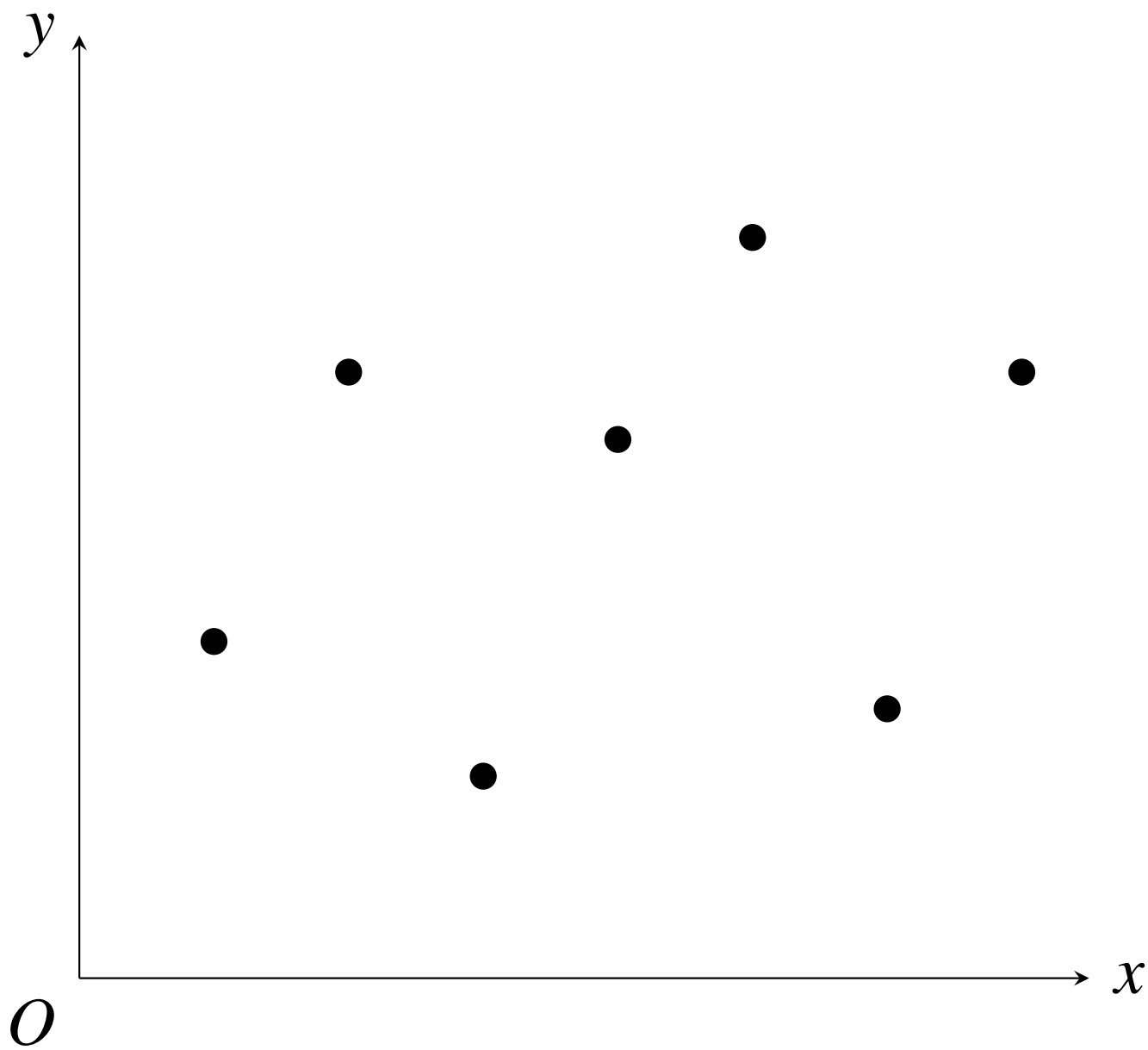
$$E(y_i \mid x_i) = \beta_0 + \beta_1 x_i.$$

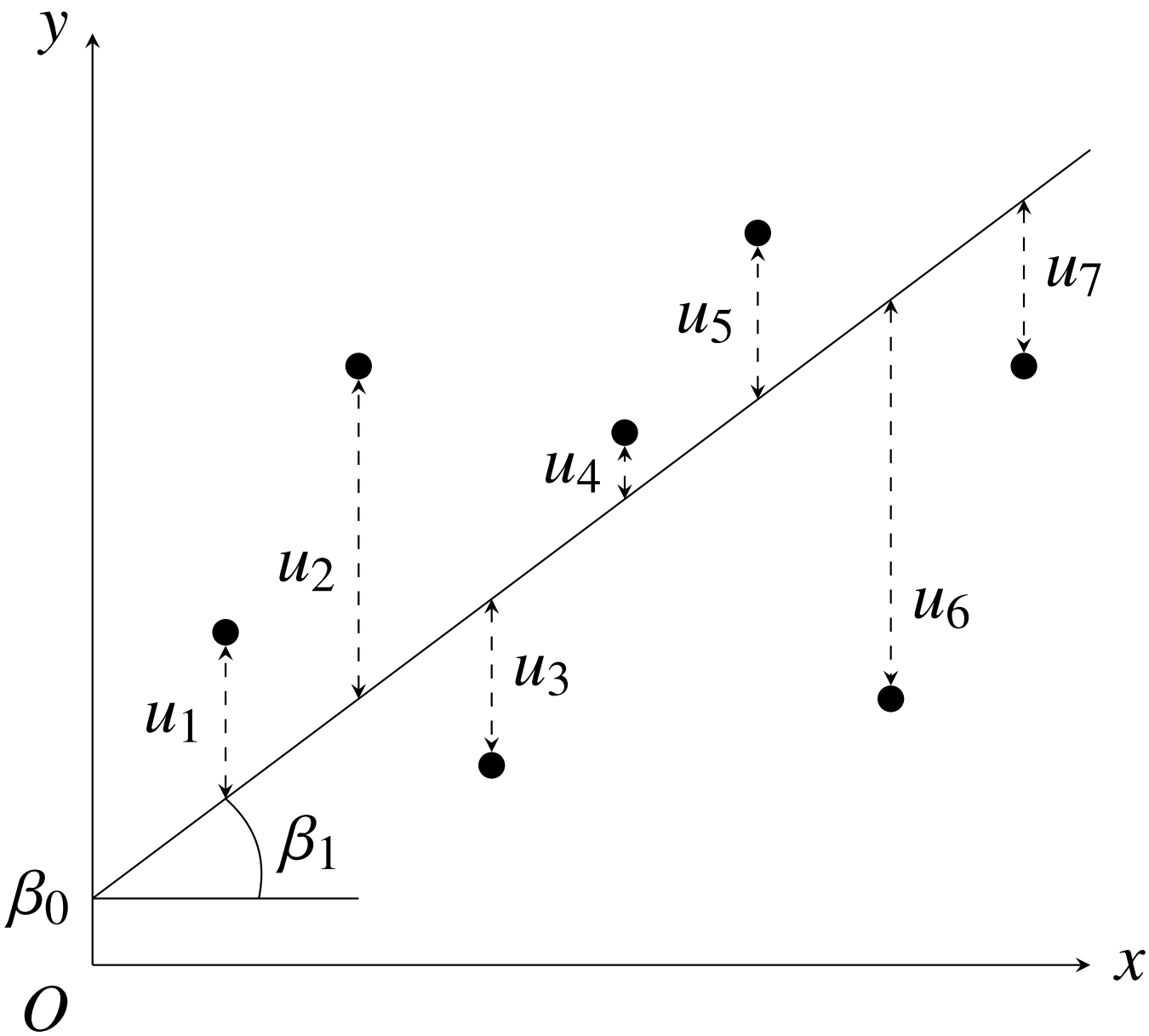
$\Rightarrow$ これは x_i が与えられたときの y_i の条件付き期待値（conditional mean）．

- $E(y_i \mid x_i)$ 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，

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

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．
- $\sum_i e_i^2$ 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，

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

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

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

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.
  - $u_i \mid x_i \sim N\left(0, \sigma^2\right).$

みかんの需要関数の推定

いま整理・加工・分析している市場別・月別データセットを用いて,
みかんの需要関数

$$q_{it} = \beta_0 + \beta_1 p_{it} + u_{it}$$

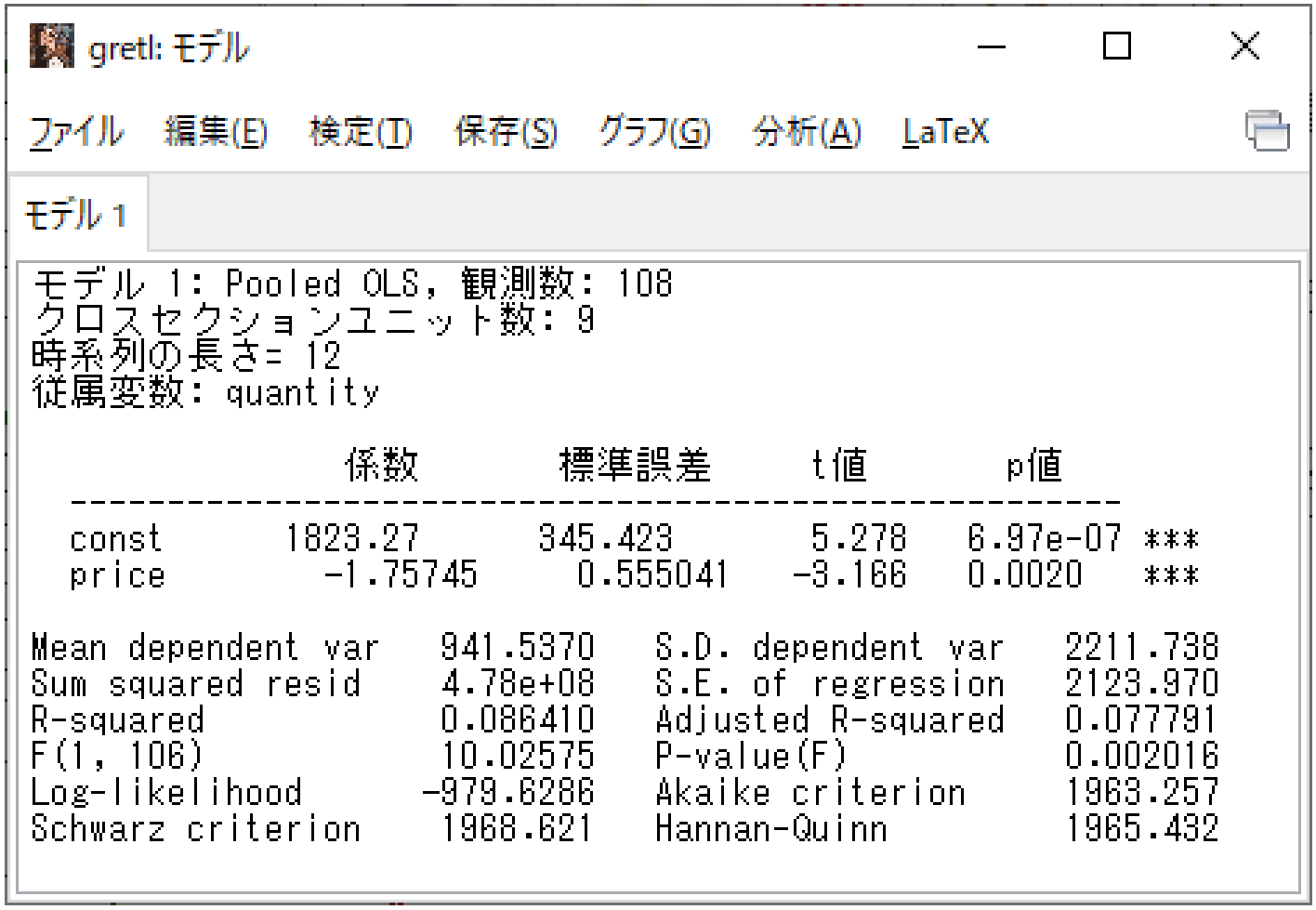
- q_{it}：取引数量（単位：t）
- p_{it}：価格（単位：円/kg）
- i：市場番号
- t：月（時点番号）

を推定する.

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. orangetokyo.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 quantity 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 quantity（みかんの取引数量）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 price 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 price（みかんの価格）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. 「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！

9. 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。
10. 「標準テキスト」を選び,「OK」をクリック。
11. 需要関数推定結果 1.txt という名前で「2020 ミクロデータ分析 2」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: 回帰係数推定値
- 標準誤差: 回帰係数の標準誤差
- 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）
- p 値: 両側 p 値
- 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，被説明変数の変動のうち，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.

$\Longrightarrow R^2 = 0$ 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.

標準誤差

- 推定量の標準偏差の推定値を標準誤差（standard error）という．
- 回帰係数の OLS 推定量 $\hat{\beta}_0$ と $\hat{\beta}_1$ の（デフォルトの）標準誤差は，それぞれ

$$\text{s.e.}\left(\hat{\beta}_0\right) = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2} \cdot \left\{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right\}},$$

$$\text{s.e.}\left(\hat{\beta}_1\right) = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2} \cdot \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

$\Rightarrow$ これらの標準誤差は，任意の i について $V(u_i \mid x_i)$ が一定（均一分散）の場合のみ正しい．

クラスター頑健標準誤差

- パネルデータのモデルでは，同じ個体について異なる時点の誤差項同士が相関する可能性がある．
  - 各個体における誤差項のクラスター（cluster）構造．
- クラスター構造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．
- gretl では，Arellano のクラスター頑健標準誤差を出力できる．
  - 「gretl: モデル推定」ダイアログボックスの，「頑健標準誤差を使用する」をチェックすればよい．データセットをパネルデータとして読み込んでいれば，その右のボタンに「Arellano」と表示される．

パネルデータを用いた実証分析では，

- 各個体における誤差項がクラスター構造をなしていることを前提としてクラスター頑健標準誤差を計算すべき．
  - 「同じ個体について異なる時点の誤差項同士が相関しない」という仮定は非現実的であることが多い．

$$\Downarrow$$

- 「ミクロデータ分析 I」で登場した，不均一分散に対して頑健な，White の頑健標準誤差は使用してはいけない．
  - 異なる時点の誤差項同士の相関がなくても，パネルデータの場合は White の頑健標準誤差を用いることに理論的正当性がない．

参考：西山慶彦・新谷元嗣・川口大司・奥井亮（2019）『計量経済学：Econometrics: Statistical Data Analysis for Empirical Economics』有斐閣, pp.232-233.

仮説検定

- $y_i = \beta_0 + \beta_1 x_i + u_i$ の y_i や x_i は様々な値をとり，観測される前はどのような値になるかが不確定（確率変数，random variable）.
- y_i や x_i を用いて計算する $\bar{y}$ や $\bar{x}$ の値も不確定.
- $y_i, \bar{y}, x_i, \bar{x}$ を用いて計算する $\hat{\beta}_0$ や $\hat{\beta}_1$ の値も不確定.

$\Longrightarrow$ 例えば回帰係数 β_1 の推定値として $\hat{\beta}_1 = -1.75745$ という値が得られても，「推定値 $\hat{\beta}_1$ は真の β_1 の値と必ずしも同じではなく，真の β_1 は 0 で，その推定値 $\hat{\beta}_1$ は様々な値をとりうる中でたまたま -1.75745 になった」可能性もある.

$\Longrightarrow$ 仮説検定（hypothesis testing）を行い，「真の β_0 や β_1 が 0 かどうか」を検証する.

gretl などの統計解析ソフトで線形回帰モデルを推定すると，各回帰係数 β_j（単回帰の場合 $j = 0, 1$）について，

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0$$

を検定するのに必要な情報が出力される．

- 回帰分析では，通常は両側検定を行う．

帰無仮説

- とりあえず「真」であると想定する仮説を帰無仮説（null hypothesis）という.
  - H_0 と書くことが多い.
  - e.g., $H_0 : \beta_1 = 0$.
  - H_0 は必ず「=」または「$\leq$ や $\geq$」を使った式.
    「$\beta_1 < 0$」を H_0 とする検定は不可能.
- まずは H_0 が「真」であると仮定し，それを「偽」とするための証拠を探す.
  - 刑事裁判における推定無罪の原則と同様.

  $\Longrightarrow$ 具体的には，検定統計値を計算する.
- 標本の関数を統計量（statistic）という.
  - e.g., 標本平均，標本分散など
- 検定に用いる統計量を検定統計量（test statistic）といい，その実現値を検定統計値という.

- 仮に H_0 が真であれば，計算した検定統計値が5%や 1%のわずかな確率でしか生じえない値になっている

$$\Downarrow$$

それを証拠として H_0 を偽と判断し，H_0 を棄却する（reject）.

- 仮に H_0 が真であれば，計算した検定統計値が小さすぎない確率で生じうる値になっている

$$\Downarrow$$

H_0 を偽とする証拠が不十分であり，偽とはいえないと判断し，H_0 を採択する（accept）.
  - 15%や 20%は「小さすぎない」.
  - 「H_0 は真」という判断ではない.

対立仮説

- H_0 が偽のときに代わりに採択する仮説を対立仮説（alternative hypothesis）という.
  - H_1 と書くことが多い.
  - e.g., $H_1 : \beta_1 \neq 0$.
  - H_1 は「$\neq, <, >$」を使った式で設定できる.

- 両側検定（two-sided test）問題の定式化：

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0$$

  - H_0 の意味は，「その説明変数は被説明変数と相関していない」
  - H_1 の意味は，「その説明変数は被説明変数と相関している」

有意水準

- H_0 が真なのに棄却することを第 1 種の誤り（type I error）という.
- H_0 を真としたときに，検定統計値が「わずかな確率でしか生じえない値」かの判断の基準となる確率，また，許容する第 1 種の誤りの確率を有意水準（significance level）という.
  - 通常は 10%，5%，1%に設定.
  - e.g., 「有意水準 5%で H_0 が棄却された」

    $\Rightarrow$ 仮に H_0 が真であれば，そんな検定統計値が出てくる確率は 5%以下に過ぎない（H_0 を偽とする証拠）ので H_0 を棄却.

    $\Rightarrow$ 言い換えると，H_0 が真のとき，「そんな検定統計値」は 5%以下の確率で出現しうる.

    $\Rightarrow$ H_0 を棄却する第 1 種の誤りを犯すことが，多くとも 5%の確率でありうる.

- H_0（係数は 0）棄却

➥「その回帰係数は統計的に有意に 0 と異なる」と判断.

  - 「その説明変数は被説明変数と統計的に有意に相関している」と解釈.
  - 定数項の検定の場合は「定数項は統計的に有意に 0 と異なる」と解釈.

- H_0（係数は 0）採択

➥「その回帰係数は 0 と異なるとは言えない」と判断.

  - 「その説明変数は被説明変数と相関しているとは言えない」と解釈.
  - 定数項の検定の場合は「定数項は統計的に有意に 0 と異なるとは言えない」と解釈.

p 値による判断

- 検定統計量（の絶対値）が実現値（検定統計値）を超える（以上になる）確率を p 値という.
  - p 値が 0.1 以下（未満）：有意水準 10%で H_0 を棄却.
  - p 値が 0.05 以下（未満）：有意水準 5%で H_0 を棄却.
  - p 値が 0.01 以下（未満）：有意水準 1%で H_0 を棄却.

$\Rightarrow$ p 値を見て，帰無仮説の採択・棄却を判断できる.

※検定統計量が連続型の確率分布（正規分布，t 分布，カイ二乗分布，F 分布など）に従う場合，「以上」と「超える」，「以下」と「未満」は区別しなくて良い.

gretl では，モデル推定結果の各説明変数の行の右端にアスタリスク（＊）が表示され，＊の個数を見れば，「有意水準何%で『回帰係数は 0』の H_0 を棄却できるか」が分かる．

- （アスタリスクなし）: 有意水準 10%でも「係数は 0」の H_0 採択．
- *: 有意水準 10%で，「係数は 0」の H_0 棄却．
- **: 有意水準 5%で，「係数は 0」の H_0 棄却．
- ***: 有意水準 1%で，「係数は 0」の H_0 棄却．

t 値による判断

定数項ありの単回帰の場合，$\beta_j = 0$ という H_0 を検定するための t 検定統計量は，

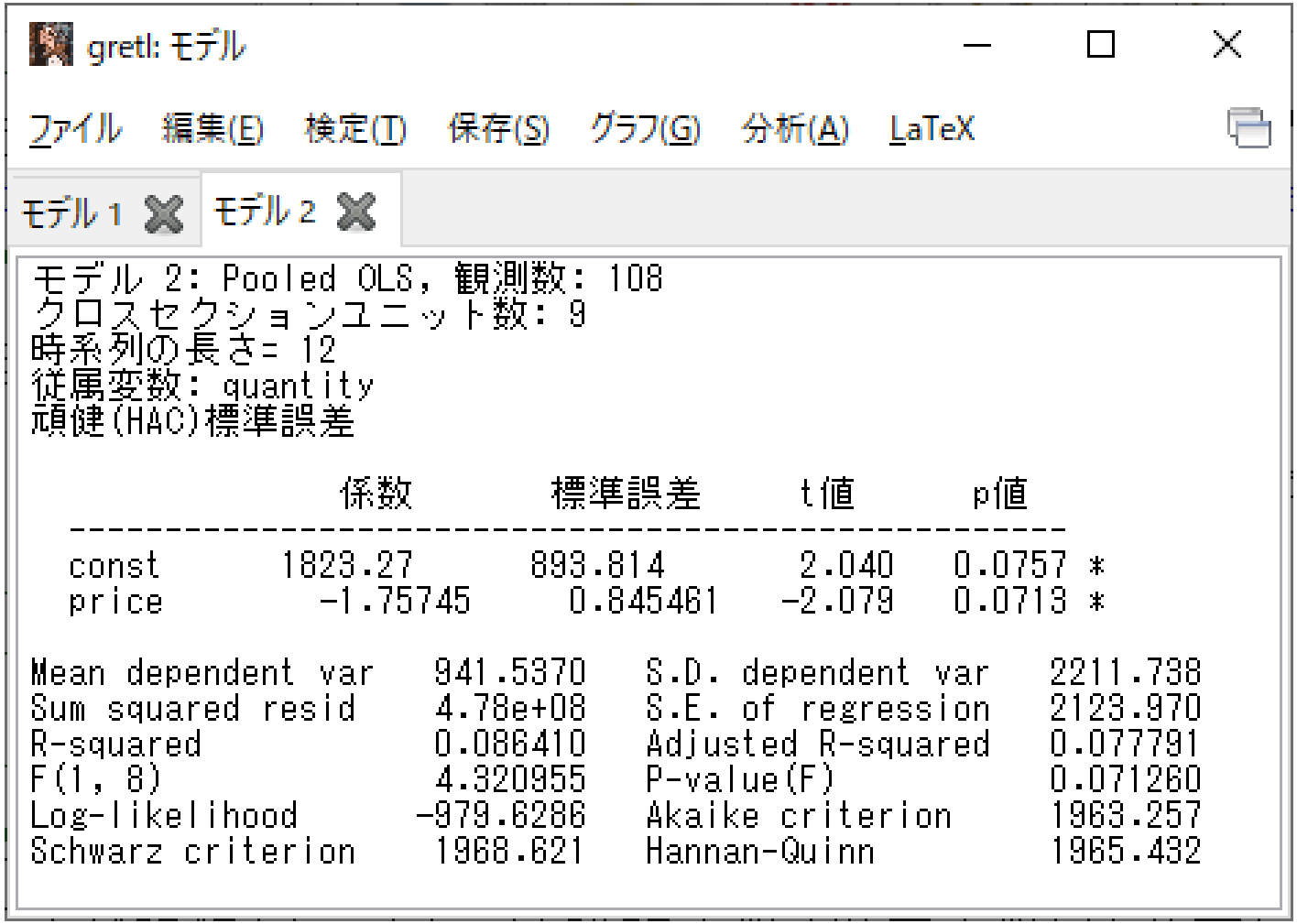
$$t = \frac{\hat{\beta}_j}{\text{s.e.}\left(\hat{\beta}_j\right)} \sim t(n-2).$$

- 観測値数が十分に大きいとき，t 値の絶対値がほぼ 2 を超えていれば，H_0 を棄却と判断（大雑把な判断）.
  ➥ 「有意水準何%で H_0 を棄却できるか」を厳密に判断するには，t 値ではなく p 値を見る.

実習 2

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 quantity 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 quantity 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 quantity（みかんの取引数量）となる．

3. 「頑健標準誤差を使用する」にチェック.
   - 誤差項のクラスター構造に対して頑健な，Arellanoの標準誤差が計算され，推定式の誤差項 u_i の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.
4. 「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．「gretl: モデル」のウィンドウはまだ閉じない！

推定結果（Arellano の頑健標準誤差）

- 価格の係数
  - −1.75745（符号は負）
  - t 値は −2.079, p 値は 0.0713
    ➥ 仮に「price の係数が 0」だとすると，−2.079 という t 値は 7.13%の確率（10%を下回る確率）でしか出てこない.
  - 有意水準 10%で，係数ゼロの H_0 棄却.
    ➥ 価格は取引数量と統計的に有意に相関している.
    ➥ みかん 1kg 当たりの価格が 1 円高くなると，取引数量は平均して 1.75745t 減少する.
    ⇒ 経済理論と整合的.

- 定数項
  - 1823.27
  - t 値は 2.04, p 値は 0.0757
    ➥ 仮に「定数項が 0」だとすると，2.04 という t 値は 7.57%の確率（10%を下回る確率）でしか出てこない.
  - 有意水準 10%で，係数ゼロの H_0 棄却.
    ➥ 定数項は統計的に有意に 0 と異なる.

- 決定係数
  - $R^2 = 0.08641.$
    ➥ 価格は取引数量の変動の約 8.6%のみ説明できている.

実習 3

1. 「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.
2. 「標準テキスト」を選び，「OK」をクリック.
3. 需要関数推定結果 2.txt という名前で「2020 ミクロデータ分析 2」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．本日の作業はここまで.