第 4 回：「ミクロデータ分析 I」の復習（4）

北村　友宏

2020 年 10 月 23 日

本日の内容

1. gretl でのダミー変数の作成
2. 重回帰モデル
3. ダミー変数を含む回帰

ダミー変数

- ある事柄が当てはまるなら 1，当てはまらないなら 0 とする変数をダミー変数（dummy variable）という．
- 各個体（市場）について，市場を表すダミー変数を作成したい．
  - e.g., 築地市場を表すダミー変数は，築地市場なら 1，それ以外なら 0 とする．

gretl での個体ダミー変数の作成

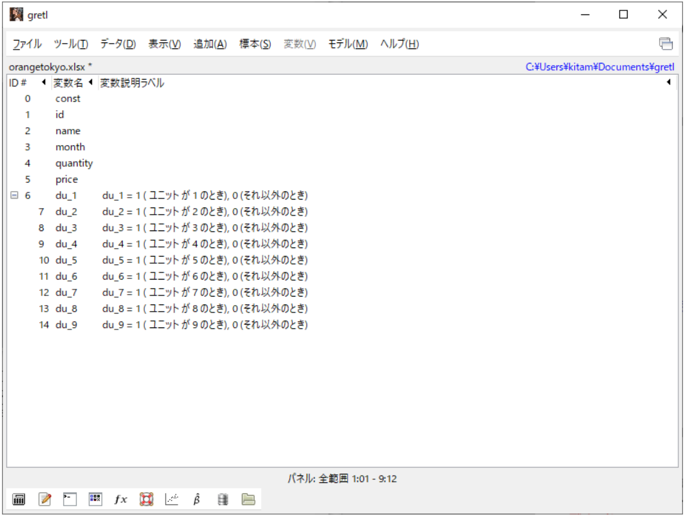
予め Excel で個体ダミー変数を作成してから gretl に読み込む方法もあるが，gretl で個体ダミー変数を作成する方法もある．

gretl で個体ダミー変数を作成するには，

- 「ユニット（グループ）インデックス変数」を「個体の識別番号の変数」に設定してパネルデータとして読み込んだ状態で，
  - この授業の実習では，第 2 回の授業で Excel データセット orangetokyo.xlsx を，「ユニット（グループ）インデックス変数」を id（個体の識別番号の変数）に設定してパネルデータとして gretl に読み込んだ．
- gretl のメニューバーから「追加」→「ユニット・ダミー」と操作すればよい．

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「追加」→「ユニット・ダミー」と操作.
   - 6 番目の変数として，「du_1」が出現し，その左端の「+」のマークをクリックすると，7 番目の変数「du_2」から 14 番目の変数「du_14」が新たに表示される.



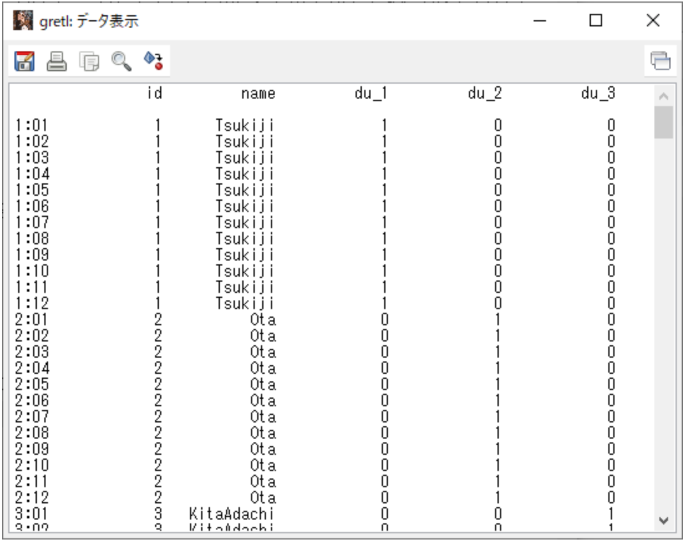
このような画面になれば成功.

5. Ctrl キーを押しながら「id」,「name」と,「du_1」から「du_9」までの計 11 個を左クリックして選択し，その上で右クリック→「データ（値）を表示」と操作すると，選んだ変数 11 個の観測値リストが新規ウィンドウにて表示される.
   - 「du_1」は，築地市場なら 1，それ以外なら 0.
   - 「du_2」は，大田市場なら 1，それ以外なら 0.
   - 「du_3」は，北足立市場なら 1，それ以外なら 0.
   - 「du_4」は，葛西市場なら 1，それ以外なら 0.
   - 「du_5」は，豊島市場なら 1，それ以外なら 0.
   - 「du_6」は，淀橋市場なら 1，それ以外なら 0.
   - 「du_7」は，世田谷市場なら 1，それ以外なら 0.
   - 「du_8」は，板橋市場なら 1，それ以外なら 0.
   - 「du_9」は，多摩ニュータウン市場なら 1，それ以外なら 0.

よって，新たに作成された変数は，

- du_1：築地市場のダミー変数
- du_2：大田市場のダミー変数
- du_3：北足立市場のダミー変数
- du_4：葛西市場のダミー変数
- du_5：豊島市場のダミー変数
- du_6：淀橋市場のダミー変数
- du_7：世田谷市場のダミー変数
- du_8：板橋市場のダミー変数
- du_9：多摩ニュータウン市場のダミー変数

と解釈できる．



	id	name	du_1	du_2	du_3
1:01	1	Tsukiji	1	0	0
1:02	1	Tsukiji	1	0	0
1:03	1	Tsukiji	1	0	0
1:04	1	Tsukiji	1	0	0
1:05	1	Tsukiji	1	0	0
1:06	1	Tsukiji	1	0	0
1:07	1	Tsukiji	1	0	0
1:08	1	Tsukiji	1	0	0
1:09	1	Tsukiji	1	0	0
1:10	1	Tsukiji	1	0	0
1:11	1	Tsukiji	1	0	0
1:12	1	Tsukiji	1	0	0
2:01	2	Ota	0	1	0
2:02	2	Ota	0	1	0
2:03	2	Ota	0	1	0
2:04	2	Ota	0	1	0
2:05	2	Ota	0	1	0
2:06	2	Ota	0	1	0
2:07	2	Ota	0	1	0
2:08	2	Ota	0	1	0
2:09	2	Ota	0	1	0
2:10	2	Ota	0	1	0
2:11	2	Ota	0	1	0
2:12	2	Ota	0	1	0
3:01	3	KitaAdachi	0	0	1

このような画面が表示されれば成功．確認したら閉じる．

6. 「ファイル」→「データを保存」と操作し，orangetokyo.gdt を上書き保存.

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．

定数項以外に説明変数が k 個ある場合，

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i,$$
$$i = 1, 2, \cdots, n.$$

各観測値の式を並べると,

$$
\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + u_1, \\
y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + u_2, \\
&\vdots \\
y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + u_n.
\end{aligned}
$$

ベクトル・行列を用いて表示すると,

$$
\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}
=
\begin{bmatrix}
1 & x_{11} & x_{12} & \cdots & x_{1k} \\
1 & x_{21} & x_{22} & \cdots & x_{2k} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & x_{n1} & x_{n2} & \cdots & x_{nk}
\end{bmatrix}
\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}
+
\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}.
$$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix},$$

$\boldsymbol{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$ とすると，重回帰モデルは次のように簡潔に表すことができる．

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

モデルを

$$\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{e},$$

と書き換え，

$$\sum_{i=1}^{n} e_i^2 = \boldsymbol{e}'\boldsymbol{e} = \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)' \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right),$$

が最小になるように OLS 推定量を求めると，

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

- $\boldsymbol{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}.$

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., 仮に全市場の条件が同じ場合の，価格が取引数量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．

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., 大田市場ダミー（大田＝ 1，それ以外＝ 0）

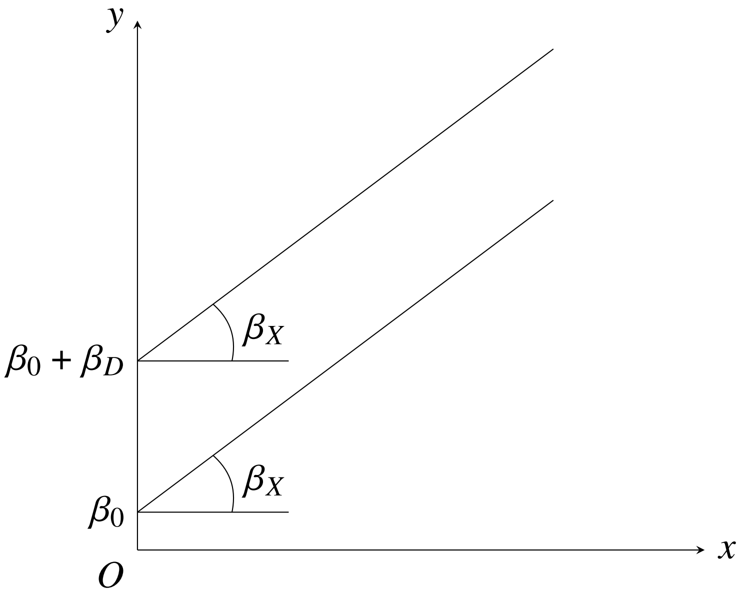
- $d_i = 0$ のとき

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

- $d_i = 1$ のとき

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

⇒ 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．
⇒ 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
β_X
$\beta_0 + \beta_D$
β_X
β_0
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., 計 9 個の市場ダミー
    「築地市場ダミー＋大田市場ダミー＋…
    ＋多摩ニュータウン市場ダミー＝ 1」
    ➥ 市場ダミーのうち 1 つを除外して，残り 8 個の市場ダミー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., 築地市場ダミーを除外して残り 8 個のダミーを説明変数に用いた場合，大田市場ダミーの係数を見れば，大田市場は築地市場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

市場名	月	築地	大田	⋯	多摩 NT	ダミー合計
築地	1	1	0	⋯	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
築地	6	1	0	⋯	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
築地	12	1	0	⋯	0	1
大田	1	0	1	⋯	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
大田	6	0	1	⋯	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
大田	12	0	1	⋯	0	1

$\Rightarrow$1 つのダミー変数と別の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，ダミー以外の説明変数の偏回帰係数の推定値，標準誤差，t 値，p 値は変わらない．
- 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．

市場ダミー変数を含む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，
市場ダミー変数を含むみかんの需要関数

$$q_{it} = \beta_0 + \beta_P p_{it} + \sum_{m=2}^{9} \beta_m d_{mi} + u_{it}$$

- q_{it}：取引数量
- p_{it}：価格
- d_{mi}：各市場ダミー
- i：市場番号
- t：月（時点番号）

を推定する．

- この定式化における仮定
  - 需要曲線の切片：各市場で異なる.
  - 需要曲線の傾き：全 9 市場で共通.

$$\Downarrow$$

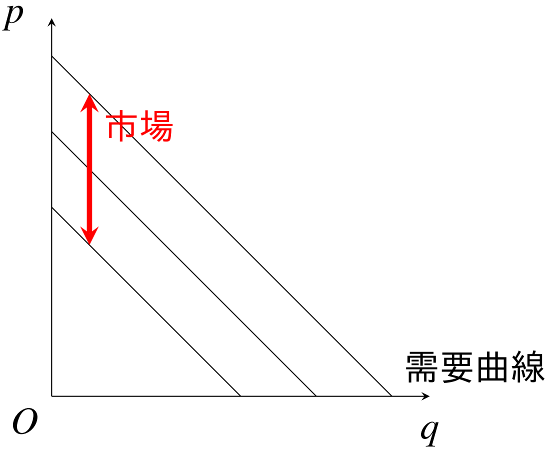
需要曲線の位置は各市場で異なる，と仮定する (消費者の数などの条件が各市場で異なるから).

- パネルデータ分析としての手法
  - 市場別・月別のパネルデータ
  - （基準となる 1 市場を除く）全市場のダミー変数を説明変数に含めて推定

$$\Downarrow$$

パネルデータ分析における Least Square Dummy Variable (LSDV) 推定を行っていることになる.

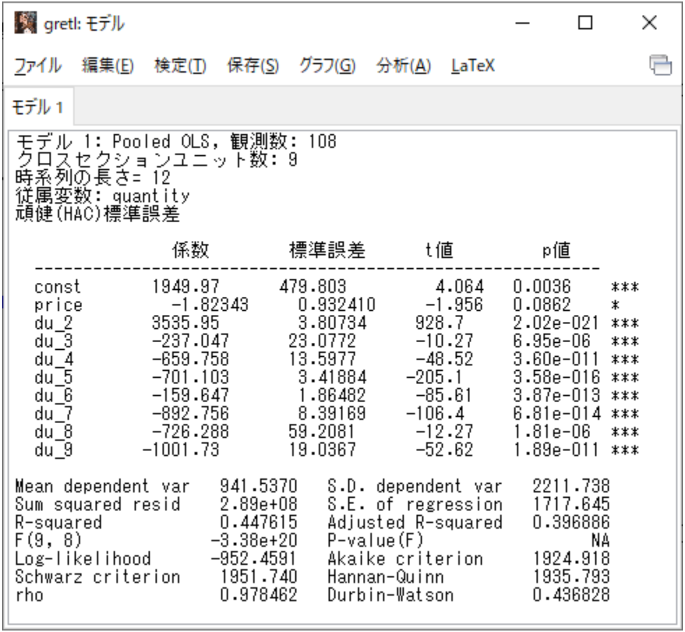
需要曲線の位置



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.
2. 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. 「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.

4. Ctrl キーを押しながら，ウィンドウ左側の変数リストにある price，du_2，du_3，du_4，du_5，du_6，du_7，du_8，du_9 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．du_1 はクリックしない．
    - 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，du_2 から du_9（築地を除く 8 市場のダミー変数）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．
5. 「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．
6. 「OK」をクリックすると，結果が表示される．




```
モデル 1: Pooled OLS, 観測数: 108
クロスセクションユニット数: 9
時系列の長さ= 12
従属変数: quantity
頑健(HAC)標準誤差

                係数          標準誤差        t値          p値
  -----------------------------------------------------------------
  const        1949.97       479.803         4.064     0.0036     ***
  price          -1.82343      0.932410     -1.956     0.0862     *
  du_2         3535.95         3.80734     928.7       2.02e-021  ***
  du_3         -237.047       23.0772      -10.27      6.95e-06   ***
  du_4         -659.758       13.5977      -48.52      3.60e-011  ***
  du_5         -701.103        3.41884    -205.1       3.58e-016  ***
  du_6         -159.647        1.86482     -85.61      3.87e-013  ***
  du_7         -892.756        8.39169    -106.4       6.81e-014  ***
  du_8         -726.288       59.2081      -12.27      1.81e-06   ***
  du_9        -1001.73        19.0367      -52.62      1.89e-011  ***

Mean dependent var   941.5370   S.D. dependent var   2211.738
Sum squared resid    2.89e+08   S.E. of regression   1717.645
R-squared            0.447615   Adjusted R-squared   0.396886
F(9, 8)             -3.38e+20   P-value(F)                 NA
Log-likelihood      -952.4591   Akaike criterion     1924.918
Schwarz criterion    1951.740   Hannan-Quinn         1935.793
rho                  0.978462   Durbin-Watson        0.436828
```

このような画面が表示されれば成功.

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．

需要関数推定結果

- 価格の係数
  - −1.82343（符号は負）
  - 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 1.82343t 減少する．
    ⇒ 経済理論と整合的．
- 定数項
  - 1,949.97
  - 有意水準 1%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．
- 自由度修正済み決定係数
  - $\bar{R}^2 = 0.396886.$
    ➥ 価格と 8 個の市場ダミー変数は取引数量の変動の約 39.7%を説明できている．

- 単回帰モデルとの価格の係数推定値の違い
  - 単回帰モデル（前回の授業で推定）では，価格の係数は -1.75745.
    ➥ 各市場の条件の違い（消費者数などの違い）による取引数量への影響が，価格の係数に反映されてしまっている.
  - 市場ダミーを含む重回帰モデルでは，価格の係数は -1.82343.
    ➥ 各市場の条件の違い（消費者数などの違い）の影響を取り除いた上で，価格が取引数量に与える影響が推定されている.
    $\Rightarrow$ 各市場の条件を一定とした状況を作り出している.
    - この分析では，
      「どの市場・時点でも，市場ダミー変数の値が 8 個とも 0（一定）」
      ＝「全 9 市場が全て築地市場（基準）と条件が同じ」という仮想的な状況.

- 各市場ダミー変数
  - 8 個の市場ダミー変数の係数は，全て有意水準 1%で係数ゼロの H_0 棄却.

市場ダミー変数の係数の解釈

推定結果に基づけば，築地市場に比べ，

- 大田市場では 3,535.95t

統計的に有意に取引数量が多く，

- 北足立市場では 237.047t
- 葛西市場では 659.758t
- 豊島市場では 701.103t
- 淀橋市場では 159.647t
- 世田谷市場では 892.756t
- 板橋市場では 726.288t
- 多摩ニュータウン市場では 1,001.73t

それぞれ統計的に有意に取引数量が少ない（有意水準 1%の両側検定）．

実習 3

1. 「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 3.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.