第 11 回：ダミー変数を含む回帰

北村　友宏

2020 年 7 月 17 日

本日の内容

1. 重回帰分析
2. 欠落変数バイアス
3. ダミー変数を含む回帰

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，

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

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

$\boldsymbol{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$ 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．

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

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

- 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., 仮に全都道府県に女性（または男性）しかいなかったときの，所得が消費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（後期「ミクロデータ分析 II」の授業で扱う）．

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

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., 男性ダミー（男性＝ 1，女性＝ 0）
  - e.g., 女性ダミー（女性＝ 1，男性＝ 0）

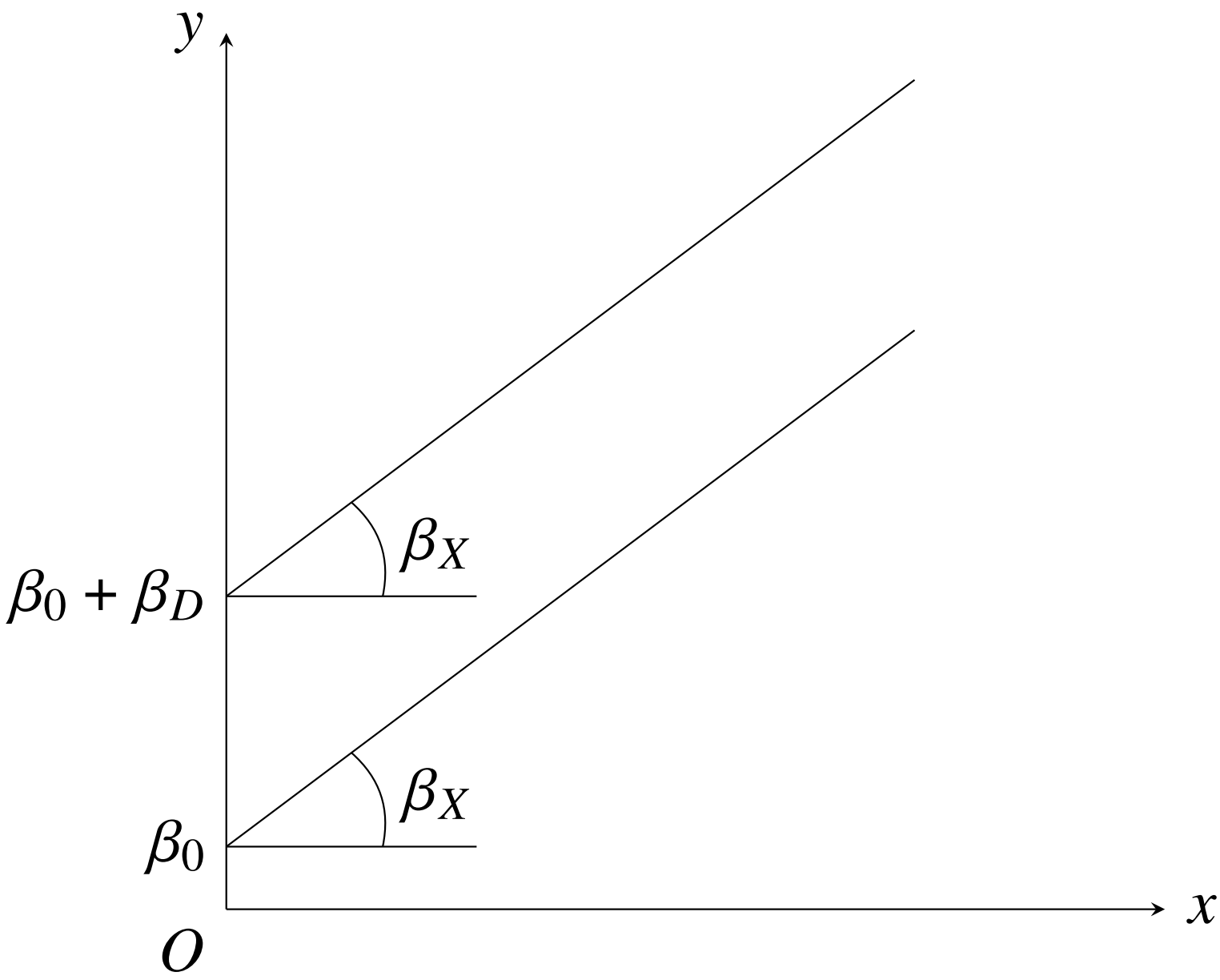
- $d_i = 0$ のとき

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

- $d_i = 1$ 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 がどの程度異なるかが分かる．


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.,「男性ダミー＋女性ダミー＝ 1」
    ➥ 男性ダミーか女性ダミーどちらか 1 つを説明変数に用いる．

$$\Downarrow$$

- 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., 男性ダミーを除外して女性ダミーを説明変数に用いた場合，女性は男性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

都道府県	男性	女性	男性 + 女性
北海道	1	0	1
青森県	1	0	1
⋮	⋮	⋮	⋮
沖縄県	1	0	1
北海道	0	1	1
青森県	0	1	1
⋮	⋮	⋮	⋮
沖縄県	0	1	1

⇓

「男性」というダミー変数と「女性」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，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．

性別ダミー変数を含む消費関数の推定

いま整理・加工・分析している都道府県別・男女別データセットを用いて，性別ダミー変数を含む消費関数

$$c_i = \beta_0 + \beta_Y y_i + \beta_D d_i + u_i \tag{1}$$

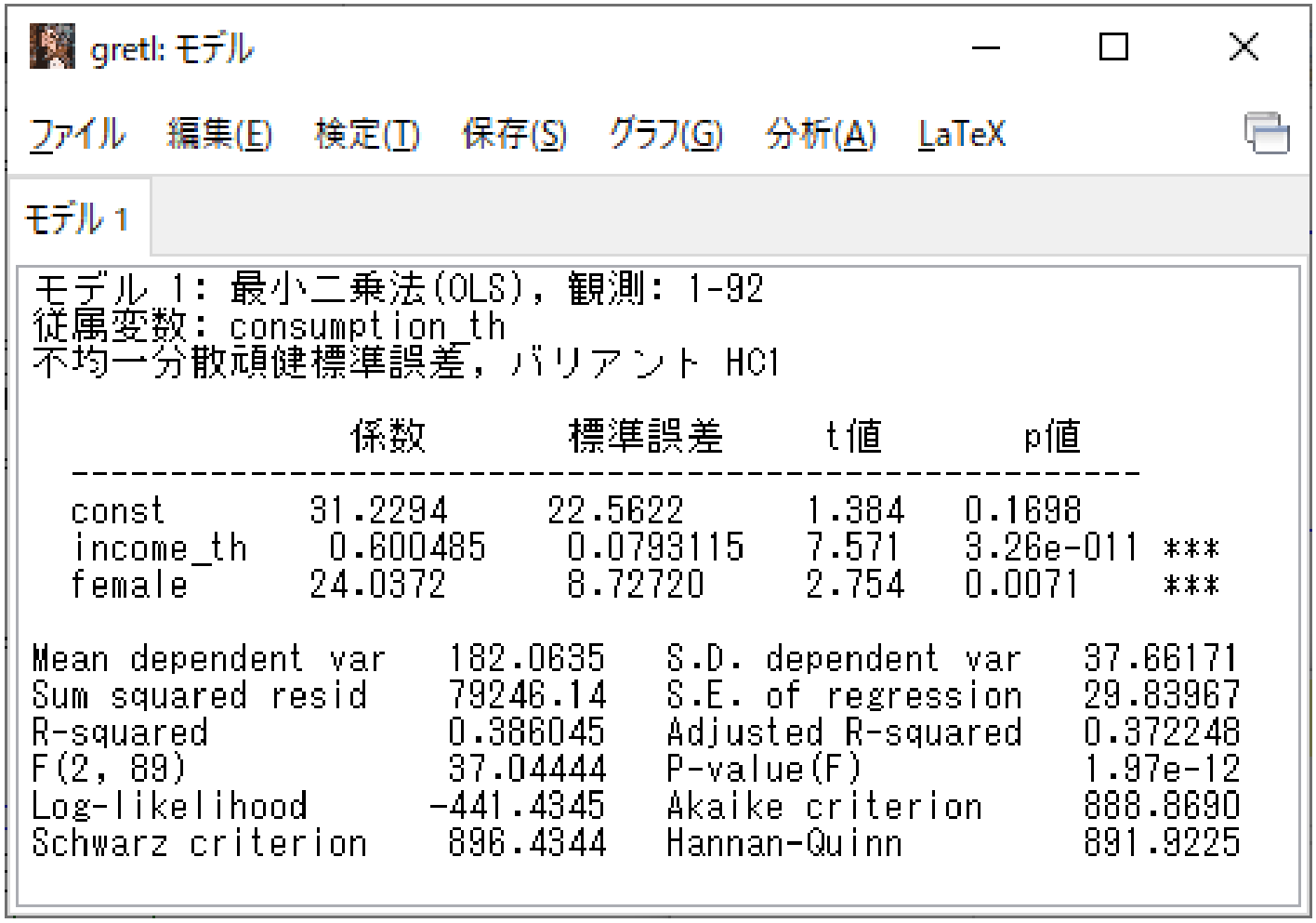
- c_i：消費支出
- y_i：可処分所得
- d_i：女性ダミー（女性＝１，男性＝０）

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. 消費 2009.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 consumption_th 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 consumption_th（千円単位の消費支出）となる.

6. ウィンドウ左側の変数リストにある income_th 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．続いてウィンドウ左側の変数リストにある female 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 income_th（千円単位の可処分所得）と female（女性ダミー）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. 「頑健標準誤差を使用する」にチェック．
   - 不均一分散に対して頑健な，White 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．
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．「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: （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，

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

$\Rightarrow$ 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．

性別ダミーを含む消費関数推定結果

- 女性ダミーの係数
  - 24.0372（符号は正）
  - 有意水準 1%で，係数ゼロの H_0 棄却.
    ➥ 所得を一定として，女性は男性に比べ統計的に有意に，消費支出額が平均して 24,037.2 円（24.0372 千円）大きい.
- 所得の係数
  - 0.600485（符号は正）
  - 有意水準 1%で係数ゼロの H_0 棄却.
    ➥ 消費と統計的に有意に相関している．性別を一定として，所得が千円高くなると消費支出額が平均して 600.485 円（0.600485 千円）高くなる.

- 定数項
  - 31.2294（符号は正）
  - 係数ゼロの H_0 採択.
    ➥ 定数項は統計的に有意に 0 と異なるとはいえない.

- 自由度修正済み決定係数
  - $\bar{R}^2 = 0.372248.$
    ➥ 所得と女性ダミーは消費の変動の約 37%を説明できている.
    ➥ ケインズ型消費関数（消費を所得のみに単回帰）の自由度修正済み決定係数 $\bar{R}^2$ は 0.313222 で，それと比較すると，モデルの当てはまりは改善されている.

レポートや論文に，消費を所得のみに回帰したモデルと，所得と女性ダミーに回帰したモデルの推定結果を載せる場合は，例えば以下のような表を載せればよい．

表 1　消費関数推定結果

	モデル（1）			モデル（2）		
	回帰係数	t 値		偏回帰係数	t 値	
所得	0.45	7.64	***	0.60	7.57	***
女性ダミー				24.04	2.75	***
定数項	78.68	5.78	***	31.23	1.38	
自由度修正済み決定係数	0.3132			0.3722		

（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 92 である．

本日の作業はここまで.