$$= \frac{1}{n-k} \sum_{i=1}^{n} (Y_i - \widehat{\beta}_1 X_{1i} - \widehat{\beta}_2 X_{2i} - \cdots - \widehat{\beta}_k X_{ki})^2$$

として表される。

このとき,

$$\begin{aligned}
&\mathrm{E}(\widehat{\beta}_j) = \beta_j, \\
&\mathrm{plim}\widehat{\beta}_j = \beta_j, \\
&\mathrm{E}(s^2) = \sigma^2, \\
&\mathrm{plim} s^2 = \sigma^2,
\end{aligned}$$

を証明することが出来る。(証明略)

分布について： $\widehat{\beta}_1, \widehat{\beta}_2, \cdots, \widehat{\beta}_k$ の分散は以下のように表される。

$$\mathrm{V}\begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \vdots \\ \widehat{\beta}_k \end{pmatrix}$$

$$
= \begin{pmatrix} \mathrm{V}(\widehat{\beta}_1) & \mathrm{Cov}(\widehat{\beta}_1,\widehat{\beta}_2) & \cdots & \mathrm{Cov}(\widehat{\beta}_1,\widehat{\beta}_k) \\ \mathrm{Cov}(\widehat{\beta}_2,\widehat{\beta}_1) & \mathrm{V}(\widehat{\beta}_2) & \cdots & \mathrm{Cov}(\widehat{\beta}_2,\widehat{\beta}_k) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(\widehat{\beta}_k,\widehat{\beta}_1) & \mathrm{Cov}(\widehat{\beta}_k,\widehat{\beta}_2) & \cdots & \mathrm{V}(\widehat{\beta}_k) \end{pmatrix}
$$

$$
= \sigma^2 \begin{pmatrix} \sum X_{1i}^2 & \sum X_{1i}X_{2i} & \cdots & \sum X_{1i}X_{ki} \\ \sum X_{1i}X_{2i} & \sum X_{2i}^2 & \cdots & \sum X_{2i}X_{ki} \\ \vdots & \vdots & \ddots & \vdots \\ \sum X_{1i}X_{ki} & \sum X_{2i}X_{ki} & \cdots & \sum X_{ki}^2 \end{pmatrix}^{-1}
$$

$\widehat{\beta}_j$ の分散 (すなわち，上の逆行列の j 番目の対角要素) を，

$$
\mathrm{V}(\widehat{\beta}_j) = \sigma^2_{\widehat{\beta}_j},
$$

として，その推定量を $s^2_{\widehat{\beta}_j}$ とする。

このとき，

$$
\widehat{\beta}_j \sim N(\beta_j, \sigma^2_{\widehat{\beta}_j}),
$$

となり，標準化すると，

$$
\frac{\widehat{\beta}_j - \beta_j}{\sigma_{\widehat{\beta}_j}} \sim N(0,1),
$$

が得られる。さらに，

$$\frac{(n-k)s^2}{\sigma^2} \sim \chi^2(n-k),$$

となり (証明略)，しかも，$\widehat{\beta}_j$ と s^2 の独立性から (証明略)，

$$\frac{\widehat{\beta}_j - \beta_j}{s_{\widehat{\beta}_j}} \sim t(n-k)$$

となる。

よって，通常の区間推定や仮説検定を行うことが出来る。

決定係数について：　また，決定係数 R^2 についても同様に表される。

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2} \\ &= 1 - \frac{\sum_{i=1}^n \widehat{u}_i^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2} \end{aligned}$$

ただし，$\widehat{Y}_i = \widehat{\beta}_1 X_{1i} + \widehat{\beta}_2 X_{2i} + \cdots + \widehat{\beta}_k X_{ki}$，$Y_i = \widehat{Y}_i + \widehat{u}_i$ である。

R^2 は，説明変数を増やすことによって，必ず大きくなる。なぜなら，説明変数が増えることによって，$\sum_{i=1}^n \widehat{u}_i^2$ が必ず減少するからである。

R^2 を基準にすると，被説明変数にとって意味のない変数でも，説明変数が多いほど，よりよいモデルということになる。この点を改善するために，自由度修正済み決定係数 $\overline{R}^2$ を用いる。

$$\overline{R}^2 = 1 - \frac{\sum_{i=1}^n \widehat{u}_i^2/(n-k)}{\sum_{i=1}^n (Y_i - \overline{Y})^2/(n-1)},$$

$\sum_{i=1}^n \widehat{u}_i^2/(n-k)$ は u_i の分散 σ^2 の不偏推定量であり，$\sum_{i=1}^n (Y_i - \overline{Y})^2/(n-1)$ は Y_i の分散の不偏推定量である。

R^2 と $\overline{R}^2$ との関係は，

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k},$$

となる。さらに，

$$\frac{1 - \overline{R}^2}{1 - R^2} = \frac{n-1}{n-k} \geq 1,$$

という関係から，$\overline{R}^2 \leq R^2$ という結果を得る。($k = 1$ のときのみに，等号が成り立つ。)

数値例： 今までと同じ数値例で，$\overline{R}^2$ を計算する。

i	Y_i	X_i	X_iY_i	X_i^2	$\widehat{Y}_i$	$\widehat{u}_i$
1	6	10	60	100	6.8	−0.8
2	9	12	108	144	8.1	0.9
3	10	14	140	196	9.4	0.6
4	10	16	160	256	10.7	−0.7
合計	$\sum Y_i$	$\sum X_i$	$\sum X_iY_i$	$\sum X_i^2$	$\sum \widehat{Y}_i$	$\sum \widehat{u}_i$
	35	52	468	696	35	0
平均	$\overline{Y}$	$\overline{X}$				
	8.75	13				

まず R^2 は，

$$
\begin{aligned}
R^2 &= 1 - \frac{\sum \widehat{u}_i^2}{\sum Y_i^2 - n\overline{Y}^2} \\
&= 1 - \frac{(-0.8)^2 + 0.9^2 + 0.6^2 + (-0.7)^2}{35 - 4 \times 8.75^2} \\
&= 1 - \frac{2.30}{10.75} \\
&= 0.786
\end{aligned}
$$

となり，$\overline{R}^2$ は，

$$
\begin{aligned}
\overline{R}^2 &= 1 - \frac{\sum \widehat{u}_i^2/(n-k)}{(\sum Y_i^2 - n\overline{Y}^2)/(n-1)} \\
&= 1 - \frac{2.30/(4-2)}{10.75/(4-1)} \\
&= 0.679
\end{aligned}
$$

となる。

注意： R^2 や $\overline{R}^2$ を比較する場合，被説明変数が同じことが必要である。被説明変数が異なる場合 (例えば，被説明変数を上昇率とするかそのままの値を用いるかによって，被説明変数が異なる)，誤差項 u_i の標準誤差で比較すべきである (標準誤差の小さいモデルを採用する)。

$\Longrightarrow$ 関数型の選択

5.3 ダミー変数について

5.3.1 異常値

データに異常値が含まれている場合，経済構造がある時期から変化した場合，ダミー変数を使う。

ダミー変数とは，0 と 1 から成る変数のことである。

例えば，データが 20 期間あるとして，9 期目のデータが，回帰直線から離れている場合 (異常値の場合) を考える。

$$D_i = \begin{cases} 0, & i \neq 9 \text{ のとき} \\ 1, & i = 9 \text{ のとき} \end{cases}$$

という変数を作り，

$$Y_i = \alpha + \delta D_i + \beta X_i + u_i$$

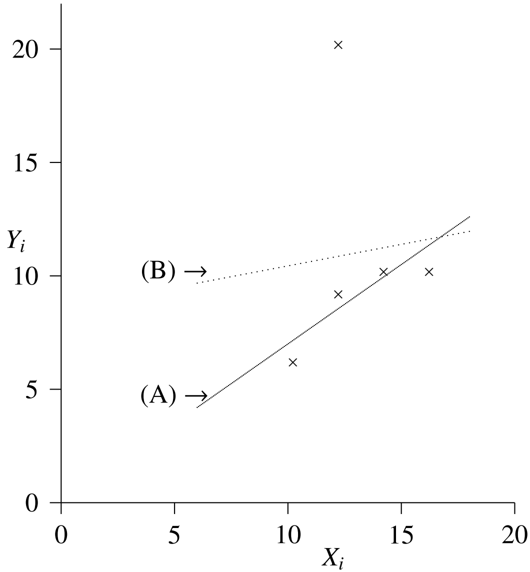
を推定する。δ の推定値 $\widehat{\delta}$ の有意性を調べることによって，異常値かどうかの検定ができる。

数値例： 今までと同様に，以下の数値例をとりあげる。

i	Y_i	X_i	D_i
1	6	10	0
2	9	12	0
3	10	14	0
4	10	16	0
5	20	12	1

第 5 期目が異常値である。

図 3： 異常値



(A) は $i = 1,2,3,4$ のデータを使って，推定した回帰直線である。(B) は $i = 1,2,3,4,5$ のデータを使って，推定した回帰直線である。(A), (B) の推定結果は以下のとおりである。

$$\text{(A):}\quad Y_i = \underset{(0.095)}{0.3} + \underset{(2.708)}{0.65}\ X_i,$$
$$R^2 = 0.786,\quad s^2 = 1.072^2,$$

$$\text{(B):}\quad Y_i = \underset{(0.49)}{8.54} + \underset{(0.14)}{0.19}\ X_i,$$
$$R^2 = 0.007,\quad s^2 = 6.09^2,$$

ただし，係数の推定値の下の括弧内は t 値を表すものとする。

このように，結果が大幅に変わる。第 5 期は異常値なので，ダミー変数を用いて，

$$Y_i = \alpha + \beta X_i + \gamma D_i + u_i,$$

として推定を行う。$i = 1,2,3,4$ について，$D_i = 0$ とし，$i = 5$ について，$D_i = 1$ とする変数である。この回帰式の意味は,

$$Y_i = \begin{cases} \alpha + \beta X_i + u_i, & i = 1,2,3,4 \text{ のとき}, \\ (\alpha + \gamma) + \beta X_i + u_i, & i = 5 \text{ のとき}, \end{cases}$$

となる。推定結果は，

$$Y_i = \underset{(0.095)}{0.3} + \underset{(2.708)}{0.65} X_i + \underset{(9.73)}{11.9} D_i,$$
$$R^2 = 0.979, \quad s^2 = 1.072^2,$$

となる。この場合，$\widehat{Y}_5 = Y_5$，すなわち，$\widehat{u}_5 = 0$ となることに注意。

5.3.2　構造変化

次に，9 期目以前と以降とで，経済構造が変化している場合を考える。

$$D_i = \begin{cases} 0, & i = 1, 2, \cdots, 8 \text{ のとき} \\ 1, & i = 9, 10, \cdots, 20 \text{ のとき} \end{cases}$$

という変数を作り，

$$Y_i = \alpha + \delta D_i + \beta X_i + u_i$$

を推定する (定数項だけが変化したと考えた場合)。または，

$$Y_i = \alpha + \delta D_i + \beta X_i + \gamma D_i X_i + u_i$$

を推定する (定数項も係数も変化)。

δ や γ の推定値の有意性を調べることによって，構造変化の検定を行うことができる。

上の例でデータを示すと，

i	Y_i	X_i	D_i	D_iX_i
1	Y_1	X_1	0	0
2	Y_2	X_2	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
8	Y_8	X_8	0	0
9	Y_9	X_9	1	X_9
10	Y_{10}	X_{10}	1	X_{10}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
20	Y_{20}	X_{20}	1	X_{20}

となる。