第 3 章　重回帰式のあてはめ

1　重回帰式の求め方

$$y=\beta_0+\beta_1 x_1+\cdots+\beta_k x_k \tag{1}$$

y ：被説明変数,　$x_1,\cdots,x_k$ ：説明変数,　β_0 ：**定数項**,

$\beta_1,\cdots,\beta_k$ ：（**偏**）**回帰係数**

β_h は，他の変数が一定で x_h だけが 1 単位増加したきの y の増分を表す.

被説明変数 y と説明変数 $x_1,\cdots,x_k$ に関してそれぞれ n（$>k$）個のデータが与えられたとき，定数項 β_0 および 回帰係数 $\beta_0,\beta_1,\cdots,\beta_k$ は，最小 2 乗法により

$$S(\beta_0,\beta_1,\cdots,\beta_k)=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_{1i}-\cdots-\beta_k x_{ki})^2 \tag{2}$$

を最小にする値として求めることができる．ここで，y_i は変数 y の第 i 番目のデータ，x_{hi} は変数 x_h（$h=1,\cdots,k$）の第 i 番目のデータを表す.

正規方程式

求める解を $b_0,b_1,\cdots,b_k$ とすると，これらは次の正規方程式をみたすことがわかる.

$$\begin{aligned}
&b_0\sum_{i=1}^{n}1+b_1\sum_{i=1}^{n}x_{1i}+\cdots\cdots+b_k\sum_{i=1}^{n}x_{ki}=\sum_{i=1}^{n}y_i\\
&b_0\sum_{i=1}^{n}x_{1i}+b_1\sum_{i=1}^{n}x_{1i}^2+\cdots\cdots+b_k\sum_{i=1}^{n}x_{1i}x_{ki}=\sum_{i=1}^{n}x_{1i}y_i\\
&\qquad\cdots\cdots\cdots\cdots\\
&b_0\sum_{i=1}^{n}x_{ki}+b_1\sum_{i=1}^{n}x_{1i}x_{ki}+\cdots\cdots+b_k\sum_{i=1}^{n}x_{ki}^2=\sum_{i=1}^{n}x_{ki}y_i
\end{aligned}$$

正規方程式が一意的な解をもつための条件

説明変数 x_1，$\cdots$，x_k および 定数 1 の間に一次従属の関係（＝ある変数が他のいくつかの変数の一次結合で表されること）がないこと.

最初の方程式に注目して，両辺を n で割ることにより，

$$\bar{y}=b_0+b_1\bar{x}_1+\cdots+b_k\bar{x}_k \tag{3}$$

が得られることがわかる．したがって，重回帰式は各変数のデータの平均の点を通ることになる.

2　重回帰のあてはまりのよさ

$$\hat{y}_i=b_0+b_1x_{1i}+\cdots+b_kx_{ki}\quad ;\text{理論値} \tag{4}$$

$$e_i=y_i-\hat{y}_i\quad ;\text{残差　＝　実績値　－　理論値} \tag{5}$$

重回帰式の性質

(a) 平均の点（$\bar{y}$，$\bar{x}_1,\cdots,\bar{x}_k$）を通る．

(b) 残差の総和は 0 となる．したがって，残差の平均も 0 となる．

(c) 残差と説明変数はすべて直交する．すなわち，

$$\sum_{i=1}^{n} e_i x_{hi} = \sum_{i=1}^{n} e_i (x_{hi} - \bar{x}_h) = 0 \qquad (h = 1,\cdots,k)$$

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

全変動　＝　回帰変動　＋　残差変動

決定係数　＝　R^2　＝　回帰変動／全変動

ただし，決定係数は次のような欠点をもっている．

〔**例題 3.1**〕　決定係数は説明変数を追加するに従って単調に増加することを示せ．

（解）説明変数が h 個のときの回帰から得られる残差変動を $RSS(h)$，決定係数を $R^2(h)$ とするとき，

$$R^2(h) = 1 - \frac{RSS(h)}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

が成立する．ここで説明変数を 1 個追加すると残差変動は $RSS(h+1)$ となるが，(2) を使うと

$$\begin{aligned} RSS(h+1) &= \min S(\beta_0,\beta_1,\cdots,\beta_h,\beta_{h+1}) \\ &\le \min S(\beta_0,\beta_1,\cdots,\beta_h,0) = RSS(h) \end{aligned}$$

となることがわかる．したがって，決定係数は $R^2(h) \le R^2(h+1)$ となる性質をもつ．

決定係数のみに頼ると，無意味な回帰をもたらす危険がある．なぜなら，被説明変数と直接に関係がないような変数を説明変数として追加しても決定係数が上がるから．

決定係数のこのような欠点を回避するために，あてはまりのよさを測る指標として

$$\overline{R}^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2 / (n-k-1)}{\sum_{i=1}^{n} (y_i - \bar{y})^2 / (n-1)}$$

を使うことが多い．ここで k は説明変数の数である．$\overline{R}^2$ を**自由度修正済決定係数**という．同一の重回帰から得られる $\overline{R}^2$ と R^2 では，前者の方が常に小さめである．

重回帰のあてはまりのよさは，実績値と理論値の間の相関の程度によっても測ることができる．このとき，

重相関係数　＝　実績値 y と理論値 $\hat{y}$ の相関係数

が定義される．理論値は回帰式から複合的に計算されたものであり，普通の相関係数と

の区別を明らかにする意味を込めて「重」という接頭語が付けられている．

重相関係数の性質

(a) 常に 0 以上，1 以下である．
(b) 決定係数の正の平方根に一致する．
(c) 単回帰においては相関係数の絶対値に等しい．

〔**例題 3.2**〕 **付録 1** のデータは，1971 年から 95 年の日本における賃金上昇率（w），完全失業率（u），物価上昇率（p）の年次データである．このとき，w を u で説明する回帰直線，および w を u と p で説明する重回帰式を求めよ．また，自由度修正済決定係数の観点から 2 つの回帰の結果を比較せよ．

（解）**図 3-2** には賃金上昇率（縦軸）と完全失業率（横軸）の散布図が示されている．これら 2 つの変数間にはトレード・オフの関係があることが見てとれる．これは**フィリップス曲線**として知られる関係である．単回帰および重回帰の結果は次の通りである．

$$w = 31.441 - 10.932u \qquad R^2 = 0.720, \overline{R}^2 = 0.707$$
$$w = 16.115 - 5.542u + 0.721p \qquad R^2 = 0.887, \overline{R}^2 = 0.877$$

(i) $\overline{R}^2$ の観点からは重回帰により説明力が上がったといえる．すなわち，賃金上昇率は単に完全失業率とのトレード・オフの関係だけでなく，インフレ要因によっても説明される．

(ii) 単回帰の場合と比較すると，定数項の値が減り，代わりに失業率の係数値が大きくなっている．これは，単回帰においては失業率の係数が物価の変動の影響も反映しているのに対して，重回帰式では物価の影響を分離していることによる．

(iii) **図 3-3**（34 ページ）は 2 つの回帰から得られる残差のプロットである．当然ながら，重回帰の残差プロットは，単回帰の場合よりもばらつきが小さくなっている．

(iv) 完全失業率については，その逆数を使った重回帰を考える方がよりあてはまりがよいことがわかる．

3 データの変換と回帰

非線形関数と線形化変換の例

(a) $y = \alpha x^{\beta}$	$Y = \log y, X = \log x$	$Y = \log\alpha + \beta X$	$(x>0, y>0, \alpha>0)$
(b) $y = e^{\alpha+\beta x}$	$Y = \log y$	$Y = \alpha + \beta x$	$(\ y>0\ )$
(c) $y = \dfrac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$	$Y = \log\left(\dfrac{y}{1-y}\right)$	$Y = \alpha + \beta x$	$(\ 0<y<1\ )$

(a) は被説明変数，説明変数ともに対数変換する例であり，$\dfrac{d\log y}{d\log x} = \dfrac{dy/y}{dx/x} = \beta$ となる

から，β は弾力性を表す．

(b) は被説明変数のみ対数変換するものである．弾力性は βx となり x に比例する．また，限界性向は βy であり y に比例する．この関数は成長に関する時系列データに適用されることが多い．

(c) の非線形関数 y は**ロジスティック関数**と呼ばれる．y がとりうる値は 0 と 1 の間であり，β が正のときには 0 から 1 へ単調に増加するから累積相対度数と同じ性質をもつ．この非線形関数の線形化変換

$$Y = \log\left(\frac{y}{1-y}\right)$$

を**ロジット変換**という．$y = 0.5$ となるときの x の値は $x = -\alpha/\beta$ であるが，ロジスティック関数が単調増加の場合には分布としての意味をもつので，この値は分布のメディアンをもたらす．

〔**例題 3.3**〕 **付録 2** のデータは，アメリカの金属産業に属する 27 社それぞれの付加価値額（y），労働投入額（x_1），資本設備額（x_2）に関するものである．このとき，**コブ＝ダグラス型生産関数**，すなわち $\log y$ を $\log x_1$ と $\log x_2$ で説明する重回帰式を求めよ．

（解）求める重回帰式は

$$\log y = 1.171 + 0.603\log x_1 + 0.376\log x_2 \qquad R^2 = 0.944, \overline{R}^2 = 0.939$$

となる．労働の弾力性と資本の弾力性の和が 1 となるとき，労働と資本の投入量を λ 倍すれば付加価値額も λ 倍となるので，付加価値額は規模に関して収穫不変となる．今の場合，弾力性の和は 0.979 である．母集団において 1 とみなすことができるかどうかの検定は第 5 章で説明する．

〔**例題 3.4**〕 **表 3-2** は世帯主の年令階級（x）ごとの持ち家率（p）と月収総額の平均（z，単位：万円）を調べたものである．持ち家率をロジット変換して，(i) x で説明する単回帰，(ii) x と z で説明する重回帰を求め，両者を比較せよ．

（解）回帰の結果は，それぞれ次のようになる．

$$y = -1.811 + 0.524x \qquad R^2 = 0.944, \overline{R}^2 = 0.933$$

$$y = -4.51 + 0.285x + 0.0347z \qquad R^2 = 0.992, \overline{R}^2 = 0.989$$

持ち家率を説明するのに，年令だけでなく所得要因も考慮した方が $\overline{R}^2$ が上がることがわかる．単回帰の結果に従えば，持ち家率のメディアンを与える x は，$x = 1.811/0.524 = 3.46$ であり，ほぼ 40 歳に対応する．**図 3-4(i)** にはロジット変換後のデータと上で得られた 2 つの回帰式が示されている．単回帰は，低，高年令層で過大，中間層で過小となっているのに対して，重回帰のあてはまりは良好である．**図 3-4(ii)** は原データと 2 つの回帰の結果から得られたロジスティック関数が図示されている．原データと単回帰，重回帰の関係は前の図と同様であることがわかる．

４　ダミー変数

回帰における説明変数としては，量的変数だけでなく，定性的な要因で変動する質的変数も使うことができる．例えば，横断面データでは，性別，学歴，職業などを表す変数，また，時系列データでは，構造変化や季節要因を表す変数を考えることができる．しかし，実際の分析では質的要因を適当な方法で数値化して，結果的には量的変数にする必要がある．このように，質的要因を量的な変数に変換したものを**ダミー変数**という．

定数項ダミー

横断面データにおいて，個人の給与 y を説明する回帰を考えよう．このとき，勤続年数 x のほかに性別要因を考慮して，男性は 1，女性は 0 をとる変数 d を導入して，賃金関数

$$y=\alpha+\beta x+\gamma d$$

を考えることができる．このとき，男性と女性の給与は，別々の賃金関数

男性：　$y=\alpha+\gamma+\beta x$

女性：　$y=\alpha+\beta x$

に従うことになり，その違いは切片に反映されることになる．切片の差 γ は初任給の差である．このような変数 d は**定数項ダミー**と呼ばれる．

係数ダミー

男女間の給与の差が，初任給だけでなく年々のベース・アップの差にもあるとき，回帰式では，その差は傾きの違いにも反映されることになる．そのためには，あらたに説明変数 $d\times x$ を導入した賃金関数

$$y=\alpha+\beta x+\gamma d+\delta(d\times x)$$

を考えることができる．このとき，男女の賃金関数は，それぞれ

男性：　$y=\alpha+\gamma+(\beta+\delta)x$

女性：　$y=\alpha+\beta x$

となり，ベース・アップの差は δ で測ることができる．この場合の変数 $d\times x$ は**係数ダミー**と呼ばれ，せいべつと勤続年数との相乗効果を表している．

交互作用ダミー

相乗効果を表すダミー変数は，一般に**交互作用ダミー**と呼ばれる．上で説明した係数ダミーもその例であるが，ここではその他の例を考えよう．今，給与を説明する要因として，勤続年数と性別のほかに学歴を加えることにしよう．学歴は性別と同様に質的要因であるが，**カテゴリー**（＝質的変数がとりうる値）は高校卒，大学卒，大学院卒の 3 通りあるものとすると，ダミー変数は 2 個必要となることがわかる．高校卒は 1，その他は 0 とする変数を l_1，大学卒は 1，その他は 0 とする変数を l_2 とすれば，賃金関数として

$$y=\alpha+\beta x+\gamma d+\delta(d\times x)+\kappa_1 l_1+\kappa_2 l_2$$

を考えることができる．さらに，この回帰を拡張して，性別と学歴，学歴と勤続年数の相乗効果を考慮した回帰は次のようになる．

$$y=\alpha+\beta x+\gamma d+\delta(d\times x)+\kappa_1 l_1+\kappa_2 l_2$$
$$+\mu_1(d\times l_1)+\mu_2(d\times l_2)+\rho_1(l_1\times x)+\rho_2(l_2\times x)$$

ここで，積の形になっている変数が相乗効果を表すダミー変数である．このとき，実際には次の 6 本の賃金関数が存在することになる．

男性，高卒：　$y=\alpha+\gamma+\kappa_1+\mu_1+(\beta+\delta+\rho_1)x$
男性，大卒：　$y=\alpha+\gamma+\kappa_2+\mu_2+(\beta+\delta+\rho_2)x$
男性，大学院卒：　$y=\alpha+\gamma+(\beta+\delta)x$
女性，高卒：　$y=\alpha+\kappa_1+(\beta+\rho_1)x$
女性，大卒：　$y=\alpha+\kappa_2+(\beta+\rho_2)x$
女性，大学院卒：　$y=\alpha+\beta x$

構造変化ダミー

時系列データにおいては，政治情勢や経済情勢の変化（例えば戦争，石油危機など），すなわち構造変化があった時点を考慮した回帰式を考えなければならない場合がある．この目的で使われるダミー変数は構造変化ダミーと呼ばれ，導入する方法として 2 通りが考えられる．

第一は，構造変化があった時点，例えば，時点 T^* だけをそれ以外の時点と区別して扱う方法である．このとき，時点 T^* で 1，その他で 0 をとるダミー変数 d^* を考えて，

$$y_t=\beta_0+\beta_1 x_{1t}+\delta d_t^* \qquad (t=1,\cdots,T^*,\cdots,T)$$

のような回帰式を作ればよい．このとき，実際には次の 2 つの回帰

$$y_t=\beta_0+\beta_1 x_{1t} \qquad (t\neq T^*)$$
$$y_t=\beta_0+\delta+\beta_1 x_{1t} \qquad (t=T^*)$$

を考えることになる．構造変化の時点が複数個で必ずしも継続的でない場合にも同様に考えることができる．

第二の方法は，構造変化があった時点を境にして異なる回帰を考えるものである．この場合には，時点 T^* までは 0，それ以後は 1 をとる変数 s を導入して，

$$y_t=\beta_0+\beta_1 x_{1t}+\delta(x_{1t}\times s_t) \qquad (t=1,\cdots,T^*,\cdots,T)$$

のような回帰式を考える．このときは，次の 2 つの回帰

$$y_t=\beta_0+\beta_1 x_{1t} \qquad (t=1,\cdots,T^*)$$
$$y_t=\beta_0+(\beta_1+\delta)x_{1t} \qquad (t=T^*+1,\cdots,T)$$

を扱うことになる．これは，係数ダミーの特殊ケースである．もちろん，この回帰式を拡張して交互作用ダミーを導入することも可能である．

季節ダミー

月次あるいは四半期の経済データは季節変動を含んでいるのが普通である．それは，1

年を周期とする気候や社会習慣が経済活動に及ぼす経済外的な影響であると考えられており，その大きさを説明するために使われるダミー変数は**季節ダミー**と呼ばれる．例えば，四半期データの場合にはカテゴリー（＝周期）が 4 であるから，導入すべきダミー変数は 3 つであり，以下のように定義することができる．

$$q_1 = \begin{matrix} 1 & \text{第 1 四半期} \\ 0 & \text{その他} \end{matrix} \qquad q_2 = \begin{matrix} 1 & \text{第 2 四半期} \\ 0 & \text{その他} \end{matrix} \qquad q_3 = \begin{matrix} 1 & \text{第 3 四半期} \\ 0 & \text{その他} \end{matrix}$$

季節ダミーをそのままの形で導入すると，定数項ダミーとしての意味合いをもつことになる．実例を考えてみよう．

〔**例題 3.5**〕　**付録 3** は 1980 年から 1995 年における日本のマクロ経済に関する四半期データである．このデータにおいて，家計貯蓄（y）を家計可処分所得（x）と季節ダミーで説明する重回帰式を求めよ．

（解）上で定義した 3 つの季節ダミー q_1, q_2, q_3 を使って，重回帰式は

$$y = 10.23 + 0.102x - 16.466q_1 - 3.721q_2 - 10.385q_3 \quad R^2 = 0.918, \overline{R}^2 = 0.913$$

となる．季節ダミーの係数は，第 4 四半期との貯蓄レベルの差を表しており，いずれも負の値をとっている．なお，季節ダミーを使わない単回帰式は

$$y = -11.14 + 0.321x \qquad R^2 = 0.468, \overline{R}^2 = 0.459$$

となり，決定係数は非常に小さく，また，限界貯蓄性向はかなり大きな値となってしまう．**図 3-5** には，家計貯蓄の実績値（実線），および単回帰に基づく理論値（点線）と季節ダミーを導入した重回帰に基づく理論値（破線）がプロットされている．季節ダミーを入れた重回帰は現実のデータの動きをよく捉えていることが見てとれる．

5　多重共線性

説明変数間の相関が非常に強い状況で起きる問題であり，重回帰式が不安定になったり，求めることすら不可能になったりする．例えば，2 個の説明変数 x_1 と x_2 があって，これらの変数の間に線形の関係 $x_2 = c + dx_1$ が成り立つものとする．このとき，重回帰式は

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 \\ &= \beta_0 + c\beta_2 + (\beta_1 + d\beta_2) x_1 \end{aligned}$$

となる．最初の等号からは 3 本の正規方程式が得られるが，実は 2 番目の等号が示すように，独立な正規方程式は 2 本しか得られないことがわかる．しかし，未知数は β_0, β_1, β_2 の 3 つである．2 本の方程式から 3 つの未知数を求めることはできない．

このように，説明変数間に従属的な線形関係が存在するために，正規方程式から回帰係数を一意的に求めることができないとき，説明変数間に**多重共線性**があるという．多重共線は 2 つの説明変数間だけでなく，3 つ以上の説明変数の間でも起こりうる問題である．

実際には，説明変数間に完全な線形関係が成立することはまれであり，厳密な意味で

の多重共線が起きることはほとんどない．しかし，ほぼ線形の関係があるような場合は，重回帰式が求められるにしても，それは非常に不安定なものとなる．

〔**例題 3.6**〕 **付録 5** の都道府県別データにおいて，行政職員数を人口と県民所得に回帰したときの重回帰式を，全都道府県を対象にした場合と，人口を 300 万人以下に限った場合の 2 通りについて求めよ．

（解）行政職員数を y（人），人口を x（千人），県民所得を z（兆円）とすると，

$y = 4123.4 - 2.35x + 1135.1z$ ：全都道府県の場合（$n = 47$）

$y = 2124.2 + 4.20x - 924.0z$ ：人口が 300 万人以下の場合（$n = 37$）

という結果が得られる．各々の重回帰式の係数の符号は正になることが予想されるが，負となるものがあり，また，2 つの重回帰式の対応する係数の符号が互いに異なっている．すなわち，重回帰式は非常に不安定である．その原因は，人口と県民所得の間の相関係数が 0.977($n = 47$)，0.978($n = 37$) と非常に高く，多重共線関係が示唆されるからである．この場合には x，z の両方を説明変数として使うべきではない．説明変数として人口のみを使った回帰は次の通りである．

$y = 1547.2 + 1.99x$ ：全都道府県（$n = 47$）

$y = 2695.2 + 1.48x$ ：人口が 300 万人以下の府県（$n = 37$）

この結果より，限界性向は 1.99（全都道府県），1.48（人口が 300 万人以下の府県）となり，これらは人口が千人増えるときの行政職員の平均的な増員数を表している．また，切片が正であることから，人口に占める行政職員の比率（平均性向）は人口の増加とともに減少し，行政職員の増加率は人口の増加率ほどではないことがわかる．

多重共線性を検出する方法としては，

(i) 説明変数間の相関係数を求める

(ii) 得られた係数の理論的な符号条件と照らし合わせる

(iii) 例題 3.6 のように，データの一部を使った重回帰の結果を，全部を使った重回帰の結果と比べてみる

ことなどがある．そして，多重共線を回避するためには，不要な変数を除去する，あるいはデータを変換するなどして，また，可能ならばデータを追加するなど，さまざまな工夫をする必要がある．