線形回帰分析

渡邊直樹

2021 年 9 月 22 日

線形回帰モデル

線形回帰（linear regression）とは，分析対象として設定された変数の散らばりや変動を他のいくつかの変数の加重和によって説明し，それらの変数（被説明変数と説明変数）の関係に関する仮説を検証するための統計手法である．回帰分析は，推定されたそれら変数の関係（モデル）における仮説の検定を通して，より説得的なモデルを探す作業を伴う．説得的なモデルが確定すれば，そのモデルに基づいて，簡便な予測を行うこともできる．

- たとえば，一国全体での当該期 t の消費 C_t と可処分所得 Y_t の間に

$$C_t = \alpha_0 + \alpha_1 Y_t \tag{1}$$

という関係を想定するならば，過去に観察された消費と可処分所得のデータ $(C_1, \ldots, C_n)$ と $(Y_1, \ldots, Y_n)$ からパラメータ α_0 と α_1 の値を推定し，その推定値を信用してよいかをいくつかの検定を行うことで検証する．式の左辺の変数を被説明変数（explained variable）といい，右辺の変数を説明変数（explanatory variable）という．

- 消費 C は，可処分所得だけではなく，当該期 t の国債発行残高 D_t や対外資産 W_t の影響も受けると想定するならば，

$$C_t = \beta_0 + \beta_1 Y_t + \beta_2 D_t + \beta_3 W_t \tag{2}$$

という式を立て，パラメータ β_0，β_1，β_2，β_3 の値を推定し，その推定値が信用するに足るものがどうかを検討する．このように，式の右辺

に複数の説明変数がある場合の回帰分析を重回帰（multiple regression）といい，説明変数が一つしかない回帰分析を単回帰（simple regression）という．線型単回帰モデルのよく知られた（実務上の）応用例の一つは資本資産価格モデル（CAPM）である．

- (1) と (2) のうち，(1) の方が妥当な式であると判断するには，β_0 と β_1 は信頼するに足る推定値であり，β_2 と β_3 の推定値はそうではないことを立証すればよい．

標準的線形回帰モデルの仮定

m 個の説明変数について，1 期から t 期までのデータがあるとする．このときの線形回帰モデルは

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_m X_{mt} + u_t$$

と書ける．ここで，u_t は誤差項であり，説明変数の加重和では説明できない Y_t の変動は確率的な変動によるものであるとみなす．また，必ず満たされなければならない仮定ではないが，各期の u_t は正規分布に従うと仮定されることが多い．（分析者が他の分布を指定しない限り，各種ソフトウェアが算出する統計量はこの仮定の下で導出されている．）

パラメータの推定値を $\hat{\beta}_k$（$k = 1, \ldots, m$）とすると，

$$e_t = Y_t - (\hat{\beta}_0 + \hat{\beta}_1 X_{1t} + \hat{\beta}_2 X_{2t} + \cdots + \hat{\beta}_m X_{mt})$$

を残差（residual）という．($\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 X_{1t} + \hat{\beta}_2 X_{2t} + \cdots + \hat{\beta}_m X_{mt}$ を理論値（当てはめ値，fitted value ともいう）という．なお，$t+1$ 期の変数が与えられたとき，$\hat{Y}_{t+1} = \hat{\beta}_0 + \hat{\beta}_1 X_{1t+1} + \hat{\beta}_2 X_{2t+1} + \cdots + \hat{\beta}_m X_{mt+1}$ を予測値（predicted value）という．）

平均 μ，分散 σ^2 の正規分布を $N(\mu, \sigma^2)$ と書く．標準的線形回帰分析における仮定は次のとおりである．

(1) u_t の期待値は 0：$(\mu =) E[u_t] = 0$

(2) u_t の分散は一定：$E[(u_t - \mu)^2] = E[u_t^2] = \sigma^2$（分散の均一性）

(3) 任意の u_t と u_s（$t \neq s$）は互いに相関しない：$E[(u_t - \mu)(u_s - \mu)] = 0$（系列相関なし）

(4) 各説明変数 X_{kt}（$k = 1, \ldots, m$）と誤差項 u_t は相関しない．

- (1) は重要な仮定ではない．$E[u_t] = \mu$ のとき，μ は定数項に反映されるだけである．
- (2) と (3) はまとめて spherical errors とも呼ばれ，これらが満たされない場合には，パラメータの推定値に対する信用の度合いを低下させる．spherical errors の有無を検定する方法はいくつも開発されており，その補正方法も確立されている．まずは，残差をプロットしてみてほしい
- (4) は非常にシリアスな仮定である．これが満たされない場合には，パラメータの推定値が偏っていないことを保証できない．この仮定が満たされるかを検定するには残差と説明変数の相関を調べる．まずは，それらの値をプロットしてみてほしい．

推定方法の評価基準

線形回帰モデルのパラメータ推定方法に対する評価基準は次のとおりである．（パラメータ β の推定値を計算する式を推定量 $\hat{\beta}$ という．推定値とは変数の具体的な値がその式に代入されたもののこと．）

1. 不偏性（unbiasedness）: 推定量の期待値はパラメータの真の値であるという性質．$E[\hat{\beta}] = \beta$.（$E[\hat{\beta}] - \beta$ をバイアスという．）
2. 効率性（efficiency）: 推定量の分散が，他の推定量と比較して，最も小さいという性質．$E[(\hat{\beta} - \beta)^2] \leq E[(\hat{\beta}' - \beta)^2]$
3. 一致性（consisitency）: サンプル数が無限に大きくなると推定量がパラメータの真の値に（確率的に）収束する性質．サンプル数が n のときの推定量を $\hat{\beta}_n$ と表すと，任意の実数 $\epsilon > 0$ について，$\lim_{n\to\infty} P(|\hat{\beta}_n - \beta| > \epsilon) = 0$

- 最小二乗法（least squares）: 残差二乗和 $\sum_t e_t^2$ を最小にするようにパラメータ β を決定する推計方法．標準的線形回帰モデルの仮定の下では，上記 3 つの評価基準をすべて満たす．

- 決定係数 R^2 と自由度調整済決定係数 $\tilde{R}^2$：データの実現値を理論値で説明できる割合を表す．線形回帰モデルの場合にのみ前述のような意味を持つことには注意が必要である．説明変数が多くなるほど，R^2 の値は大きくなる傾向があるので，データ解析においては，それを割り引いた $\tilde{R}^2$ を参照する．ただし，線形回帰モデル自身の「性能」を表すものではないことには注意が必要．

- t 検定：各パラメータの推定値 $\hat{\beta}_k$ に対する信用の度合いを検定する．多くのソフトウェアにおいて，帰無仮説は $H_0 : \hat{\beta}_k = 0$

- F 検定:設定された線形回帰モデルに意味があるかどうかを検定する．多くのソフトウェアにおいて，帰無仮説は $H_0 : \hat{\beta}_1 = \cdots = \hat{\beta}_m = 0$. （定数項 β_0 を除く.）決定係数と自由度調整済決定係数は便利な数値ではあるが，それらではモデルの当てはまりを検定することはできない．そこで，F 検定を行う．

重回帰分析では次の 2 点に注意して分析を進める必要がある．

1. 多重共線性（multicollinearity）：説明変数間に相関があるとき，それらの変数を区別し難くなる．そのような変数のパラメータの推定量に対する信用の度合いは低下する．(t 検定における帰無仮説を棄却し難くなる.)

2. 誤った定式化：必要な説明変数が線形回帰モデルから抜け落ちた場合には，パラメータの推定量は偏りを持つ．不必要な説明変数が加えられた場合には，その他の説明変数のパラメータの推定量の分散が大きくなり，推定値に対する信用の度合いが低下する．(t 検定における帰無仮説を棄却し難くなる.)

多重共線性や誤った定式化を考慮して，説明変数の選択を行わなければならない．いくつかの変数の組み合わせを試してみて，赤池情報量基準（AIC）の値が低い線形回帰モデルを選択することが多いが，第 1 節末尾で言及した (1) と (2) のモデル選択の背後に想定する論理そのものの検証を行うことがデータ解析の目的であれば，たとえ，AIC が低くても，論理的に必要な説明変数を線形回帰モデルから落としてはならない．

（練習問題）表 1 にまとめられている練習用データにおいて，$L_t = \beta_0 + \beta_1 W_t + \beta_2 X_t + u_t$ と $L_t = \beta_0 + \beta_1 W_{t-1} + \beta_2 X_t + u_t$ では，どちらの線形回帰モデルがより妥当（説得的）だといえるか？生産者の利潤最大化問題における理論上の符号条件は $\beta_1 < 0$，かつ，$\beta_2 > 0$ である．

表 1: 練習用データ：11 年を基準に指数化してある．

年 t	雇用者数 L_t	実質賃金 W_t	生産量 X_t
1	93	60	67
2	96	69	69
3	96	81	74
4	97	84	85
5	97	81	81
6	95	89	72
7	95	96	80
8	97	102	84
9	97	113	89
10	99	111	96
11	100	100	100
12	102	104	101
13	103	107	101
14	103	113	105
15	104	118	116
16	106	123	122

Excel のアドインに組み込まれている分析ツール「回帰分析」には各係数の推定値の p 値は表示されるが，AIC の値や sperical error の有無を検定するための数値は計算されず，当然ながら，表示もされない．easy R には，それらを表示するオプションがついている．ここでは，練習問題で提示されているモデル j の赤池情報量基準（AIC_j）を次の式に従って計算してみよう．

$$\mathrm{AIC}_j = T \ln(\frac{\sum e_t^2}{T}) + 2k_j. \tag{3}$$

ここで，T は t の個数（つまり，$T = 16$），k_j はモデル j において推定される係数の数であり，ln は自然対数を表している．前述のように，e_t は t 期における残差である．Excel の分析ツール「回帰分析」には残差を計算す

るオプションはついており，該当するチェックボックスにチェックを入れることで計算結果が出力される．それを使って残差自乗和を計算し，自然対数を取るとよい．セルに $= 16 * \mathrm{LN}(\text{モデル } j \text{ の残差自乗和}/16) + 2 * k_j$ と入力すると，AIC_j が計算される．

練習問題では，たとえば，2 変数のうち，どちらか 1 つを説明変数から外して推定してみてもよい．また，16 年分あるデータを 2 分割してみて，16 年分のデータで推計するよりも「好ましい」結果が出る場合，分割した年を境にして経済の構造変化が起こったと考えることもできる．サンプルサイズは極めて小さいが，色々試してみよう．

予測

変数間の関係を表すモデルとデータが与えられたとき，被説明変数が将来とる値を予測する問題を設定してみよう．ここでは，m 個の説明変数について，線形モデル

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_m X_{mt} + u_t$$

各説明変数 X_m について，1 期から t 期までのデータ（$X_{m1}, \ldots, X_{mt}$）と X_{mt+1}，および，1 期から t 期までの被説明変数のデータ（$Y_1, \ldots, Y_t$）が手許にあるとして，予測値 $\hat{Y}_{t+1}$ をどのように計算するかという問題に限定して，予測に関する簡単な説明を付しておく．

具体的には，次の手続きをとる．

(1) t 期までのデータを用いて，回帰分析を行い，パラメータの推定値 $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m$ を求める．

(2) それら推定値を用いて，$\hat{Y}_{t+1} = \hat{\beta}_0 + \hat{\beta}_1 X_{1t+1} + \hat{\beta}_2 X_{2t+1} + \cdots + \hat{\beta}_m X_{mt+1}$ を求める．

(3) 予測誤差 $e_{t+1} = Y_{t+1} - \hat{Y}_{t+1}$ の標準偏差を求め，予測域を定める．

予測域は $\hat{Y}_{t+1}$ が 95%ほどの確率でその領域に値をとるように設定されることが多い．大凡の目安としては，予測域の上限値と下限値を

$$\hat{Y}_{t+1} \pm 2 \times (\text{予測誤差の標準偏差}) \tag{4}$$

と目算すると便利である．標準的線形回帰分析の仮定の下での単回帰の場合，予測誤差の分散は

$$\sigma^2[1+\frac{1}{t}+\frac{(X_{t+1}-\bar{X})^2}{\sum_1^t(X_t-\bar{X})^2}] \tag{5}$$

で与えられる．ここで，$\bar{X}$ は説明変数 X の平均を表す．

(5) から判るように，X_{t+1} が $\bar{X}$ から大きく離れた値をとるほど予測誤差の分散は大きくなる．予測誤差の標準偏差とはその分散の正の平方根である．また，予測誤差の分散は標本数が大きくなるほど小さくなる．

推定方法の評価基準と同様に，予測方法の評価基準もいくつかあり，上記手続きで求めた予測値は次の 2 つの性質を持つことが知られている．ここで，t 期の予測量を，一般に，

$$\check{Y}_{t+1} = E(Y_{t+1} : \text{t 期において利用可能な情報})$$

とし， 予測誤差を $e_{t+1} = Y_{t+1} - \check{Y}_{t+1}$ と定義する．線形回帰を使って予測を行う場合は $\check{Y}_{t+1} = \hat{Y}_{t+1}$ となっている．

1. 不偏性（unbiasedness）: 予測誤差の期待値はパラメータの真の値であるという性質．つまり，予測誤差は平均的にはゼロ．$E[e_{t+1}] = 0$.

2. 効率性（efficiency）: 予測量の分散が，他の予測量と比較して，最も小さいという性質．他の任意の予測量による予測の誤差を e^*_{t+1} とすると，$E[(e^2_{t+1}] \leq E[e^{*2}_{t+1}]$.