多変量ゼミ §1.12回帰式による予測値の区間推定 (重回帰の場合)

髙見澤 真央

ここでは，単回帰の流れをもとに，重回帰における目的変数の予測値の区間推定を行う．

　おおまかな流れとして，

　(1)重回帰モデルによる目的変数の期待値および分散を計算する

　(2) (1)で得られた結果から，信頼区間をわりだし，区間推定を行う

　このように進めていく．

　また，前節の§1.11回帰係数の推定と検定(重回帰の場合)の範囲で求めた式を参照する場合は，式番号に☆のマークを付けて表す．

1. 回帰式による予測値の区間推定(重回帰の場合)

1.1. 予測値の期待値と分散

　説明変数がある特定の値をとるときの目的変数の期待値は，回帰式を用いて，

によって推定(予測)される．このとき，の期待値は，

|  |  |  |
| --- | --- | --- |
|  |  | |
|  |  | |
|  |  | |
|  | (☆式(8),(11) より) | |
|  |  |  |

となり，予測値は に対する不定推定値であることがわかる．

　また，の分散は，

|  |  |  |
| --- | --- | --- |
|  |  | |
|  |  | |
|  | (式(1) より) | |
|  |  | |
|  |  | |
|  |  | |
|  |  | |
|  |  | |
|  |  | |
|  | (☆式(10) , ☆式(12)  ☆式(13) ) | |
|  |  | |
|  |  | |
|  | () | |
|  |  |  |

　前節において，回帰係数と定数項は，正規分布に従うがわかった．

|  |
| --- |
| いずれも正規分布に従う変数の1次式で表される ⇒ 回帰係数と定数項も正規分布に従う |

の1次式である式(1) もまた，正規分布に従うことがわかる．ここで，を標準化すると，

|  |  |
| --- | --- |
|  |  |
|  |  |

となる．は標準正規分布に従う．ただし，このとき未知の誤差分散を含んでおり，これを不偏推定値 で置き換えて得られる統計量は，

　ここで，

と置くとする．このは，単回帰における

この式の分母の(=標準偏差で標準化した平均からの距離の2乗)を，多変量の場合に一般化したものに相当し，点と重心との間のマハラノビスの汎距離と呼ばれる．

|  |  |
| --- | --- |
|  | ざっくり マハラノビスの距離とは・・・  　左の図において，赤い点のデータ群があるときに，  青い点と緑の点はどちらの点がよりデータ群から離れているかを判断したい．  データの平均からのユークリッド距離は，ほぼ同じであるが，  点の散らばりを見るとと，緑の点の方があきらかに赤い点のデータ群から離れていることがわかる．  正しく判断するには，データの散らばり具合も考慮した「データ群からの距離」を考える必要があり，これを実現しているのが  マハラノビスの距離である．  　主に判別分析や異常検知の分野で使われている． |

　式(5)を用いると，式(4)は次のようになる．

このとき， は自由度の分布に従う．

信頼区間は，平均して100回中100()回 母集団の を含むことが保証された範囲のこと．

　これから，に対する信頼率の信頼区間は，

すなわち，

このように求められる．式(7)をみると，の小さいところ，すなわち重心の近くでは信頼区間の幅が小さく，逆に が大きく，重心から遠く離れたところでは，信頼区間の幅が大きくなることがわかる．

1.2. 目的変数ｙの信頼区間と区間推定

　次に，説明変数の特定の値に対して観測される目的変数の値の信頼区間について考える．は，これまでに観測されているやこれらにもとづくとは独立に，平均，分散の正規分布に従うため，の期待値と分散は，

|  |  |  |
| --- | --- | --- |
|  |  | |
|  | (式(2) より) | |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  | |
|  |  | |
|  |  | |
|  |  | |
|  | （とは無相関のため，） | |
|  |  | |
|  | (式(3) より) | |
|  |  |  |

となる．これを用いて，を標準化し，を不偏推定値 で置き換えると，

を得る．は自由度の分布に従う．

これより，に対応して観測される目的変数に対する信頼率の信頼区間は，

すなわち，

このように与えられる．

2. 説明変数の選択

　これまでの議論では，回帰モデルに含まれる説明変数は定められたものとして計算してきた．しかし，実際の現象を分析する場合には，目的変数に影響を及ぼす可能性のある数多の変数の中から，次のようなことを考慮し，変数を取りあげるのが一般的である．

　1. 回帰モデルに無駄な変数(真の回帰係数が0であるような変数)が含まれる場合，

　　誤差分散の推定値の自由度が小さくなり，回帰係数や予測値の推定制度が悪くなる．

　2. 必要な変数(真の回帰係数が0でない変数)が回帰モデルから外れている場合，

　　回帰係数の推定値や目的変数の予測値が偏りを持ち，誤差分散の推定値は過大評価になる．

　3. 説明変数の中に互いに相関が高い変数が含まれる場合，

　　分散共分散行列の行列式が0に近くなるため，逆行列の要素が大きくなり，

　　回帰係数の推定精度が悪くなる．特にある説明変数と残りの変数の重相関係数が のとき，

の行列式は0となり，逆行列が存在しない．そのため回帰係数を求めることができない．

これを多重共線性の問題という．

|  |
| --- |
| 多重共線性とは・・・  　重回帰分析を行ったとき，互いに関連性の高い説明変数が存在すると，解析上の計算が不安定となり，回帰式の精度が極端に悪くなる現象．またの名をマルチコ現象と呼ぶ． |

2.1. 総あたり法

　個の説明変数の候補の中から，個の変数で考えられるすべての組合せ 通りの回帰モデルを検討する方法．最も単純な方法ではあるが，候補となる変数の個数が多くなると，組合せの数が急速に大きくなり，計算時間が膨大になる．

2.2. 前進選択法

　説明変数が1つも含まれない状態からスタートして，次のような手順で変数を1つずつ増加させる．

(ⅰ) 目的変数との単相関が最大(言いかえると，1つずつ順に変数を採用してみて回帰式を計算したとき，回帰係数検定のための の絶対値または値が最大)の変数を選び，回帰係数が0であるという仮説の検定を行い，仮説が棄却されなければ，どの変数もモデルに含めない．仮説が棄却されれば，この変数をモデルにとりこみ次のステップ(ⅱ)へ進む．

(ⅱ) 既に入っている変数に加えて残りの変数を1つずつ順に採用してみて，偏相関係数が最大(回帰係数検定のための の絶対値または値が最大)の変数を選ぶ．選ばれた変数に対する回帰係数が0であるという仮説の検定を行い，仮説が棄却されなければ終了．仮説が棄却されれば，変数を取り込んで次のステップ(ⅲ)へ進む．

(ⅲ) 回帰式を計算する．もしモデルにすべての変数が含まれていれば終了．

そうでなければステップ(ⅱ)へ戻る．

2.3. 後退消去法

　説明変数の候補がすべて含まれた状態からスタートして，次のような手順で変数を1つずつ減少させる．

(ⅰ) モデルに含まれる各変数に対する回帰係数検定のための または 値を計算し，絶対値が最小となる変数を選ぶ．回帰係数が0であるという仮説が棄却されなければ，その変数を落として次のステップ(ⅱ)へ進む．棄却されれば終了．

(ⅱ) もしモデルに含まれる変数がなくなっていれば終了．そうでなければ回帰式を計算しなおして，

ステップ(ⅰ)へ戻る．

2.4. 逐次法

　前進選択法では，1度モデルに含まれた変数が途中で落とされることはなかった．この点を改良し，次のような手順で変数を増減させる．

(ⅰ) 目的変数との単相関が最大の変数を選ぶ．選ばれた変数に対する回帰係数が0であるという仮説の検定を行い，棄却されなければどの変数も回帰モデルに含めない．棄却されれば，この変数をとりこんで次のステップ(ⅱ)へ進む．

(ⅱ) 既に入っている変数に加えて残りの変数を1つずつ順に採用してみて，偏相関係数が最大の変数を選ぶ．選ばれた変数に対する回帰係数が0であるという仮説の検定を行い，仮説が棄却されなければ終了．仮説が棄却されれば，変数を取り込んで次のステップ(ⅲ)へ進む．

(ⅲ) 回帰式を計算し，各変数について回帰係数検定を行い，値が最小になる変数について，仮説が棄却されなければ，その変数を落とす．

(ⅳ) すべての変数がとりこまれていれば終了．そうでなければステップ(ⅱ)に戻る．

　上の4つの各方法において，回帰係数の検定は次のように行う．

　個の変数を含むモデルでの変数に対する回帰係数が0という仮説の検定は，

前節の式(20)

で，とおき，

の値を求めて，自由度の分布の限界値と比較し，ならば仮説を棄却，

ならば仮説を採択する．

　また，分布と分布の関係，

「あるが自由度の分布に従うとき，は自由度の分布に従う」ことから，

の値を求めて，自由度の分布の限界値と比較し，ならば仮説を棄却する，ことで同じ結果が得られる．

補足 多重共線性

　説明変数が2つの回帰モデル を例に，多重共線性の問題について考える．

回帰係数を求める式は，§1.5の線形重回帰の範囲から，

である．この式の分母，すなわち，分散共分散行列の行列式は，

|  |  |  |
| --- | --- | --- |
| (式(14)の分母) |  | |
|  |  | |
|  |  | (15) |

となる．このとき は，との相関係数である．

との相関が高い，すなわち相関係数 がに近い値をとるとき，分散共分散行列の行列式は，0に近くなり，回帰係数は式(14)の分子の行列式の変化に大きく影響されることがわかる．変数値の変化によって，回帰係数の推定値が大きく変わるため，係数推定値の分散が大きくなり，推定結果の信頼性がおちる．また，特に相関係数が (線形従属)のとき，分散共分散行列の行列式は0となり，それを分母にもつ式(14)は計算ができず，そもそも回帰係数を求めることができない．

　重回帰分析では，候補となる説明変数の間に相関がないことを確認し，相関が見られた場合にはその説明変数を外すことで，多重共線性の問題を回避する必要がある．