異文化言語教育評価ＩＡ　　平成 26 年 7 月 2 日
第８章　回帰分析

報告者：H.W.

§８－１　回帰分析とは

回帰分析：変数間の因果関係や方向性を想定して、１つまたは複数の独立変数から従属変数の予測の大きさ（説明率）を検討する場合に用いる。

単回帰分析：１つの独立変数から予測
重回帰分析：複数の独立変数から予測

例）アパートを借りる場合の条件（独立変数）から、アパートの家賃（従属変数）を予測することができる。

独立変数		従属変数
駅からの距離		
築年数	予測 ⇒	アパートの家賃
部屋の広さ		

・パラメトリック検定だが、従属変数および独立変数ともに、間隔尺度または比率尺度に、２値の名義データ（例：木造・鉄筋）を加えて重回帰分析を行うことができる。

８－１－１　単回帰分析と単回帰式

１つの独立変数から従属変数を予測することを単回帰分析という。

単回帰式：$Y' = b_1 X + b_0$（b_0：定数）

Y ：従属変数。目的変数ともよぶ（例：期末テストの得点）
Y'：予測値。回帰式（モデル）で予測された値（例：i 番目の観測値；図 8.1 の Y_i'）
X ：独立変数。説明変数または予測変数ともよぶ。予測に用いられる変数。
b_1 ：回帰係数。独立変数が従属変数に与える影響力，直線の傾き
b_0 ：定数。回帰直線が縦軸と交わる点，切片

例）期末テストの予測点＝（b_1×勉強量）＋b_0（普段の勉強でとれる得点，常に変わらない値）

回帰係数（b_1）および定数（b_0）の値は、回帰直線が実際のデータに最もよく適合するように計算される。その計算式を導く方法として最小２乗法がある。

残差平方和＝Σ（観測値－予測値）^{2}　➡　残差平方和が最小になるよう計算

８－１－２　重回帰分析と重回帰式

複数の独立変数から従属変数を予測することを重回帰分析という。
重回帰式は単回帰式の応用で、複数の独立変数が式に追加された直線モデル。

重回帰式：　$Y' = \underbrace{b_1X_1+b_2X_2+b_3X_3}_{\text{変動する部分}} + \underbrace{b_0}_{\text{定数}}$

Y'：予測値。重回帰式（モデル）で予測された値
X_1, X_2, X_3：独立変数
b_1, b_2, b_3：各独立変数の偏回帰係数

例）家賃＝（b_1×駅からの距離＋b_2×築年数＋b_3×部屋の広さ）＋b_0（定数,基準家賃）
「駅からの距離」の偏回帰係数（b_1）が−1500 だったとすると：
➡1 km 離れたアパートは家賃が 1,500 円（−1500×1）安くなる
10 km 離れたアパートは家賃が 15,000 円（−1500×10）安くなる

各独立変数（b_1, b_2, b_3）の影響力を比較する場合：
従属変数および全ての独立変数の平均を 0、分散を 1 に標準化する。そのときの偏回帰係数を**標準（化）偏回帰係数**とよび、1 に近いほど影響力が大きいといえる。

●偏回帰係数の解釈（留意点）

①重回帰式は、従属変数（Y',家賃）を変化させる部分（$b_1X_1+b_2X_2+b_3X_3$）と変化させない部分（b_0）から成り立っている。
➡定数部分（b_0,基準家賃）の割合が、変動する部分（$b_1X_1+b_2X_2+b_3X_3$）よりかなり大きい場合、いくら偏回帰係数（b_1, b_2, b_3）が大きくても、独立変数（駅からの距離、築年数、部屋の広さ）の従属変数全体（家賃）に及ぼす影響は小さくなる。

例）定数（b_0）が 30 万円の場合：
偏回帰係数（b_1, b_2, b_3）によって家賃を変動させることができる割合は全体として小さくなり、どのような条件でも比較的家賃が高くなる。

②偏回帰係数の大きさは従属変数（家賃）と独立変数（駅からの距離、築年数、部屋の広さ）との因果関係の強さまでは示していない。
➡偏回帰係数を因果関係の強弱として解釈する場合は、理論的な根拠が必要になる。

③標準偏回帰係数は、単独では従属変数（家賃）に対して大きな影響力を持つ独立変数であっても、他の独立変数の従属変数への予測力に影響され、どのような変数を投入するかによって、標準偏回帰係数が小さい値になったり、ときに負の値になったりすることがある。

➡従属変数と独立変数の関係だけでなく、独立変数同士の相関や多重共線性の問題（p161）などを考慮に入れて解釈する必要がある。

８－１－３　重相関係数と決定係数

重回帰分析によって示されるもの

1. 偏回帰係数・b_1, b_2, b_3：個々の独立変数の予測力を示したもの
2. 重相関係数・R：独立変数全体から得られた従属変数との相関を示したもの
   ※単回帰分析の場合:算出されるRは従属変数との相関係数となり、標準回帰係数と同じになる。
3. 決定係数・R^2：独立変数全体でどのくらい従属変数を説明しているかを示したもの
   ※決定係数は回帰式のあてはまりの良さを表しており、１に近いほどあてはまりがよいといえる。

決定係数の算出にあたり３種類の分散が算出される

① SS_T：平均と個々の観測値の差の２乗を足し合わせた全平方和（全変動）のこと。
従属変数の平均が観測値のモデルとしてどの程度、適切であるかを表す。（傾き０の場合の分散）

② SS_R：残差平方和のこと。
観測値が回帰直線からどの程度ずれているかを表す。

③ SS_M：SS_T から SS_R を引いた平方和で、回帰直線による変動のこと。
従属変数の平均値より回帰直線を予測に使うことで、どの程度予測がよくなったかを示す。
➡決定変数（R^2）は全変動（SS_T）に対する回帰直線による変動（SS_M）の割合を示す。

●F 検定（分散分析）

決定係数が有効かどうか、つまり、従属変数の平均を使うより回帰式のほうが観測値のあてはまりがよいかの検定結果が図 8.3（p160）のように表示される。
➡有意でない場合はモデルを使って予測する意義がないことを示す。

§８－２　回帰分析を行う際の注意点

８－２－１　回帰分析の前提

特に重回帰分析を行う場合には、関わってくる前提が多く注意が必要。

（1） サンプルサイズと質

・信頼性のある決定係数を得るために　50＋8k（k＝独立変数の数）のサンプルが必要
・各独立変数の有意性を検定するに　104＋k（k＝独立変数の数）のサンプルが必要
・誤差の少ない信頼性の高いデータであることが望ましい。

高い説明率（独立変数がどのくらい従属変数を説明できているか）が期待できるのであればサンプルサイズは 80 で十分だが、高い説明率が期待できないのであれば、サンプルサイズは 200 以上必要となる。

※ 制限付きデータであれば、正確な分析はできない。

（2） 多重共線性

・独立変数間で非常に高い相関がある場合、回帰式の信頼性が低くなる場合がある。このような独立変数間の関係から生じる問題を**多重共線性**という。独立変数間の相関係数が .80 以上であれば（.90 以上であれば必ず）多重共線性を疑う必要がある。さらに、許容度と VIF の指標で多重共線性が発生していないか診断する。

例）「駅からの距離」、「駅からの所要時間」が独立変数の場合、よく似た変数であるため高い相関があると考えられる。➡相関係数をチェック＋許容度と VIF の指標をチェック

① 許容度：ある独立変数を従属変数として、他の独立変数群から予測した場合に得られる決定係数の値を、1 から引くことで求められます。引いた値が.10 以下のときに多重共線性が生じていると判断される（p169；図 8.12）。

② VIF の指標：許容度の逆数（VIF=1/許容度）で、10 以上であると多重共線性が発生しているとされる。10 未満であっても 10 に近い値になっていれば注意が必要（p169；図 8.12）。

<u>多重共線性が生じていると判断できる場合の対策</u>

➡相関の高い 2 つの独立変数のうち、1 つを分析から外す。
➡相関の高い 2 つの独立変数の平均値あるいは因子得点（p199）などの合成得点を使う。

（3） 外れ値

・回帰直線は外れ値に大きく影響されるので、データに外れ値が含まれていないかを事前に調べる必要がある。

① 残差：各データの残差を標準値（z 得点）に変換し、その標準偏差± *2*SD または± *3* SD 以上の値の割合を調べる（p167；図 8.8，p170；図 8.14）。

・±2SD 以上の値をとるデータの数が全体の 5%以内であれば問題ない
・±2.5SD 以上の値をとるデータの数が全体の 1%以内であれば問題ない
・0.1%以下の確率で起こる±3.3SD 以上の値は検討の余地がある（残差）

② クックの距離：データが回帰式全体に与える影響を示す指標であり、この値が 1 以上であれば、問題があると考えられる（p168；図 8.9，p170；図 8.15）。

③ てこ比：各ケースにおける複数の変数データが全体の平均からどの程度ずれているかを示す指標で、0 から 1 までの値をとる。この値が平均てこ比（p162；式 8.10）で求められる値の 3 倍以上の値をとるケースは問題がある（p168；図 8.9，p170；図 8.15）。

④ マハラノビス距離：複数の独立変数における各データの平均が交差する重心と各ケースのデータの距離を示す指標であり、この値が大きいデータは外れ値である可能性がある。マハラノビスの表を参考にカットポイントを決める（p168；図 8.9，p170；図 8.15）。
※てこ比と同じような外れ値が検出されるので、いずれか一方の指標を用いればよい。

（4）残差の独立性、正規性、等分散性、線形性

・回帰分析では、残差に関して①独立性、②正規性、③等分散性、④線形性の 4 つが満たされているという前提がある。

①独立性：どの独立変数の残差間にも相関がないという前提。ダービン・ワトソン検定（p167；図 8.8，p168；図 8.11）で調べることができる。0 から 4 までの値をとり、この値が 2 に近いほどよいと考えられる。

②正規性：残差の散布図やヒストグラムを作成し、データが正規分布していることを確認する（p168；図 8.10，p171；図 8.16）。

前提が満たされない場合

➡データの変換

➡線形回帰分析から非線形回帰のロジスティック回帰分析への切り替え

③等分散性：独立変数がどの値のときも残差分散は同じである（等質性がある）必要がある。つまり、残差は予測した回帰直線に沿って同じように散らばっていることが望ましい。（p168；図 8.10，p171；図 8.17）

かなり異なっている場合（p163；図 8.4（b,d））

➡不等分散性があり、その他の要因が予測に影響していると考えられる

④線形性：線形回帰分析の場合、残差は予測値（Y'）と線形関係にある必要がある（p163；図 8.4（a））。これは、標準残差と標準予測値の関係を散布図にして調べることができる（p168；図 8.10，p171；図 8.18）。

線形の関係が成り立っていない場合（p163；図 8.4（c,d））

➡線形回帰分析から非線形回帰のロジスティック回帰分析への切り替え

８－２－２　投入法

・重回帰式が有意でも、モデルに投入された全ての独立変数が有意とは限らないので、重回帰式の有意性と各独立変数の有意性とは別に考える必要がある。
・独立変数をどの順序で重回帰式に投入するかによって各独立変数の有意性および偏回帰係数が変化する。
➡目的に合った投入法を用いて解釈することが大切

（１）　強制投入法：全ての独立変数を一度に投入する方法
- 全ての独立変数でどの程度従属変数を説明することができるのか、また、従属変数の予測における各独立変数の独自の寄与がどの程度であるかを調べるのに使用する。
- 関係のない独立変数であっても、分析に投入されると決定係数は大きくなるため、本当に重要な変数を過少評価することにつながる。そのため、理論や仮説にもとづいて慎重に選んだ独立変数のみを投入するようにする。
- 予測に有効な独立変数のみで再分析を行い、回帰式および決定係数を算出することもある。

（２）　階層的投入法（階層的回帰分析）：理論や仮説に基づいて、独立変数を一つずつ投入して行く方法
- 従属変数の予測に重要とされる変数から投入することで、理論的に優先する独立変数の説明率を調べるために使用する。
- 強制投入法の分析後に行うことで、各独立変数の説明率が変化する過程を確認できる。

（３）　ステップワイズ投入法（統計的回帰分析）：統計的に最も予測率が高いと考えられる変数から順に自動的に投入される方法
- 適合度が最良の重回帰式を調べる際に使用する。
- 独立変数を投入するごとに除去すべき変数がないかを分析できる。
- 統計的な根拠に基づいて投入されるため、投入された独立変数が理論にかなっているかは別途判断する必要がある。

その他

変数増加法：ステップワイズ法と同様に独立変数を順に投入していく方法
- 独立変数を投入するごとに、除去すべき変数がないかは分析できない。

変数減少法：最初に全ての独立変数を投入し、予測への寄与が小さい独立変数から順に変数を抜いていく方法

以上の投入法を視覚的に表したのが図 8.5 のベン図

（1）強制投入法

偏回帰係数　IV_1…a，IV_2…c，IV_3…e

重相関 R … a,b,c,d,e 全てが反映

決定変数 R^2 … a,b,c,d,e 全てが反映

（2）階層的投入法

Step1　偏回帰係数　IV_1…a,b

重相関 R … a,b が反映

決定変数 R^2 … a,b が反映

Step2　偏回帰係数　IV_1…a，IV_2…c,d

重相関 R … a,b,c,d が反映

決定変数 R^2 … a,b,c,d が反映

Step3　偏回帰係数　IV_1…a，IV_2…c，IV_3…e

重相関 R … a,b,c,d,e が反映

決定変数 R^2 … a,b,c,d,e が反映

（3）ステップワイズ法　　（d＋e）の説明部分が最も大きいとすると・・・

Step1　偏回帰係数　IV_3…d,e

重相関 R … d,e が反映

決定変数 R^2 … d,e が反映

Step2　偏回帰係数　IV_3…d,e，IV_1…a,b

重相関 R … a,b,d,e が反映

決定変数 R^2 … a,b,d,e が反映

Step3　偏回帰係数　IV_3…e，IV_1…a，IV_2…c

重相関 R … a,b,c,d,e が反映

決定変数 R^2 … a,b,c,d,e が反映

誤植：p 159 の図 8.2 は、②と③の図が入れ替わっている。

補足：

1．分析の手順…外れ値が観察された場合は、まず外れ値を除外して分析をし直すことが大事。外れ値を除外することで、偏回帰係数や有意確率などが変わる可能性がある。

2．投入法…理論に基づいた独立変数の投入が難しい場合は、ステップワイズ法を利用すると便利である。しかし、ステップワイズ法のみから結論を導くと解釈を誤る場合があるので、強制投入法など他の方法の結果も参考にしながら分析をすすめたほうがよい。