第 8 章前半（pp.155-166）

# 第 8 章　回帰分析

## 8-1　回帰分析とは

一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

単回帰分析：一つの従属変数（$Y$）を一つの独立変数（$X$）から予測する。

例）「アパートの家賃」を「駅からの距離」という条件から予測する。

重回帰分析：一つの従属変数（$Y$）複数の独立変数（$X_1, X_2, X_3 \cdots, X_n$）

例）「アパートの家賃」を「駅からの距離」、「築年数」、「部屋の広さ」から予測する。

### 8-1-1 単回帰分析と単回帰式

（図 8.1 参照）回帰分析の場合は線形モデル（linear model）を立て、それぞれの観測値から最も近くなる直線を求める。

（式 8.1）観測値（observed）＝予測値（model）+残差（deviation/residual）

（式 8.2）残差平方和（$SS_R$）＝Σ（観測値－予測値）$^2$

（式 8.3）単回帰式：$Y'=b_1X+b_0$

求めた直線：回帰直線

予測値と観測値のずれ：残差/誤差

回帰直線と $Y$ 軸が交わる点：切片（$b_0$）：$X$ が 0 のときの $Y$ の値

回帰係数（$b_1$）：回帰直線の傾き

### 8-1-2　重回帰分析と重回帰式

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。それぞれの偏回帰係数を求める。

（式 8.6）重回帰式：$Y'=b_1X_1+b_2X_2+b_3X_3+b_0$

Y'：　予測値

$X_1, X_2, X_3$：独立変数

$b_1, b_2, b_3$：各独立変数の偏回帰係数

・標準（化）偏回帰係数（$\beta$）が 1 に近いほど影響力が大きい

**・偏回帰係数の解釈**

① 重回帰式は従属変数を変化させる部分 $b_1X_1+b_2X_2+b_3X_3$ と変化させない定数部分 $b_0$ から

成り立っている。➔　定数部分が大きければ、独立変数の偏回帰係数が大きくとも、従属変数全体に及ぼす影響は小さくなる。

② 偏回帰係数の大きさは従属変数と独立変数の因果関係の強さまでは示していない。

③ 標準偏回帰係数は単独では従属変数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る。➔単純に各標準偏回帰係数の値を従属変数への影響の大きさの違いとみなすことは危険。

### 8-1-3　重相関係数と決定係数

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表す。

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す。1 に近いほど回帰式のあてはまりがよいことを表す。単回帰分析の場合は $R$ は標準回帰係数と同じ値になる。

（式 8.8）　決定係数 $R^2$ = 全変動 $SS_M$/ 回帰直線による変動 $SS_T$

① $SS_T$(total sum of square)：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（前変動）

② $SS_R$(residual sum of square)：残差平方和。観測値が回帰直線からどの程度ずれているかを表す

③ $SS_M$(model sum of square)：$SS_T$から $SS_R$を引いた平方和。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使うことで、どの程度予測がよくなったかを示す。

**・F 検定(分散分析)**

モデルの分散 $SS_M$を自由度で割った平均平方和（$MS_M$）が、モデルの誤差分散（$SS_R$）を自由度で割った平均誤差分差（$MS_R$）よりどの程度大きいかを分散比（$F$値）として算出する。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

**（1）　サンプルサイズと質**

比較的多くのサンプルが必要（決定係数にとっては 50+8k（k ＝独立変数の数）、各独立変数の有意性にとっては 104+k）。誤差の少ない信頼性の高いデータであることが必要。

**（2）多重共線性**

独立変数間の関係から生じる問題。独立変数間で非常に高い相関がある場合、本来は関係ないはずの独立変数が従属変数の予測に貢献していると現れてしまったりする。相関係数.80 以上であれば多重共線性を疑う必要がある。例）「駅からの距離」と「駅からの

所要時間」が独立変数の場合、高い相関が関係あると考えられる。①許容度と②VIF の指標で多重共線性が発生していないか診断する。

① 許容度：ある独立変数を従属変数として他の独立変数軍から予測した場合に得られる決定係数の値を、1 を引くことで求められる。この値が .10 以下のときに多重共線性が生じていると判断される。

② VIF：許容度の逆数（VIF=1/許容度）で 10 以上あると多重共線性が生じていると判断される。

もし多重共線性が発生している場合は、相関の高い２つの独立変数のうち１つを分析から外すか、相関の高い２つの独立変数の平均値あるいは因子得点などの合成得点を使う、などの対策をとる。

### （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、その標準偏差の±2$SD$ または 3$SD$ 以上の値の割合を調べる。

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

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

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

### （４） 残差の独立性、正当性、等分散性、線形性

データの残差について以下の４つが満たされているという前提が必要

① 残差の独立性：どの独立変数の残差間にも相関がない。ダービン・ワトソン検定で値が 1 以下あるいは 3 以上の場合に問題がある。

② 残差の正規性：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か確認する。この前提が満たされない場合はデータを変換したりすることを検討する必要がある。

③ 残差の等分散性：独立変数がどの値のときも残差分散は同じ（等質性がある）必要がある。

④ 残差の線形性：残差は予測値（$Y'$）と線形関係にある必要がある。標準残差と票っユん予測値の関係を散布図にして調べることができる。

## 8-2-2 投入法

（１） 強制投入法：全ての独立変数を一度に投入して従属変数の予測を行う方法

（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 1 つずつ投入していく方法。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。

（3） ステップワイズ法 / 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

図 8.5 ベン図

（1）強制投入法 : $IV_1$ 偏回帰係数では a、$IV_2$ 偏回帰係数では c、$IV_3$ 偏回帰係数では e が反映される。それぞれの係数が重なっている部分 b と d も従属変数の説明率に寄与しているので、a から e すべてが重相関（$R$）より決定係数（$R_2$）に反映される。

→ 解釈の差異には、偏回帰係数と同時に表示することができる、もともとの相関、偏相関、部分相関も参考にする。

（2）階層的投入法 : $IV_1$、$IV_2$、$IV_3$ の順に投入された各時点での決定係数および偏回帰係数が分かる。$IV_2$ が投入された場合、決定係数に関しては c + d のみが R および R2 の増加分となり、R2 変化量に反映される。b に関しては $IV_1$ が投入された時点で R および R2 に反映されているので変化しない。ただし、$IV_1$ の偏回帰係数は $IV_2$ と重なるために小さくなり、a の部分のみが $IV_1$ の偏回帰係数としては反映され、$IV_2$ の偏回帰係数では c + d のみが反映される。最終的に $IV_3$ が投入されると $R_2$ として e の部分が加算されるので、R と $R_2$ は強制投入法の場合と同じ値になる。また偏回帰係数もそれぞれ独自説明部分のみになるので、強制投入法の場合と同じ値になる。

（3）ステップワイズ法 : 階層的投入法と同様だが、変数が投入される順番が従属変数への寄与部分が大きい順になる。

8-3-1 強制投入法
8-3-2 出力結果の見かた（強制投入法）
8-3-3 ステップワイズ法
8-3-4 出力結果の見かた（ステップワイズ法）
8-3-5 論文への記載
8-4 ダミー変数を使った回帰分析
8-4-1 2 値の名義尺度
8-4-2 ダミー変数の作成
8-4-3 ダミー変数を含んだ階層的回帰分析
8-4-4 論文への記載