

第8章 回帰分析

- 変数間の因果関係を予測する

回帰分析とは

回帰分析：変数間にそこまで因果関係がない場合でも、変数間に方向性を想定し、独立変数による従属変数の予測の大きさを検討する分析。



(例) アパートを借りる条件 (独立変数) からアパートの家賃 (従属変数) を予測する。

回帰分析とは

- 回帰分析は大きく2種類に分けられる。

単回帰分析 (single regression) : 独立変数が1つ

重回帰分析 (multiple regression) : 独立変数が複数

※回帰分析はパラメトリック検定であるため間隔尺度・比率尺度データを扱うが、名義尺度データをダミー変数に変換して重回帰分析を行うこともできる。

単回帰分析と単回帰式

- 回帰分析等において独立変数から従属変数の予測値を求める場合には、観測値に当てはまるモデルを立て、そのモデルの当てはまりを検討する。
- 回帰分析では線形モデル (linear model) を立て、それぞれの観測値から最も近くなる直線を、最小2乗法 (method of least squares) を用いて求める。

【回帰直線の求め方】

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

(式8.2) 残差平方和 (SSR) = $\sum (\text{観測値} - \text{予測値})^2$

【回帰直線の求め方】

(式8.3) 単回帰式 : $Y' = b_1X + b_0$

(式8.4) 期末テストの予測点 = $(b_1 \times \text{勉強量}) + b_0$

(式8.5) i 番目の測定値 : $Y + (b_1X_i + b_0) + \varepsilon_i$

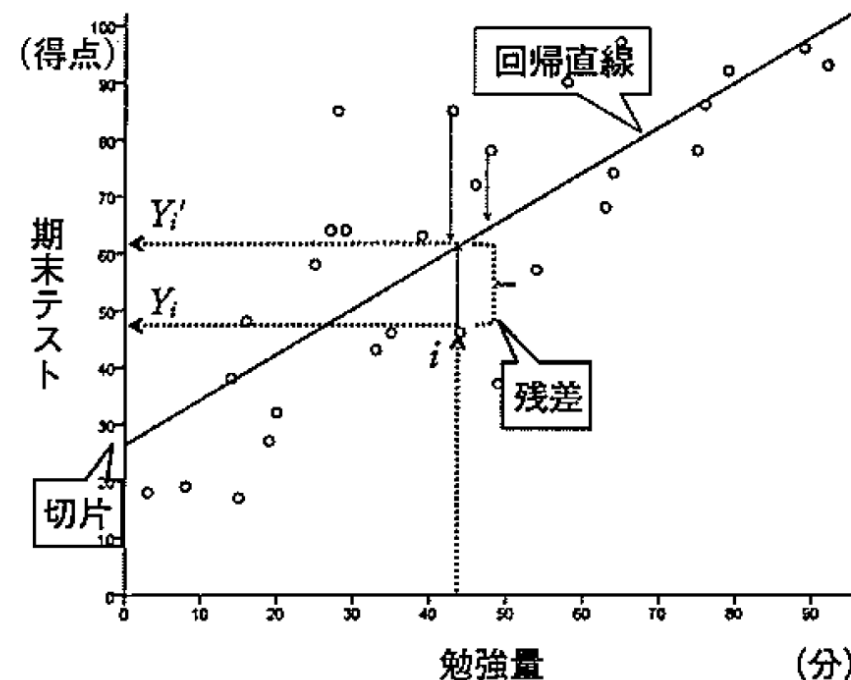


図 8.1 回帰直線

Y : 従属変数 Y' : 予測値 Y_i : 従属変数の i 番目の観測値 X : 独立変数

b_1 : **回帰係数** (regression coefficients)

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** (regression constant)

回帰直線が縦軸と交わる点、切片。

ε_i : 残差 (誤差)

重回帰分析と重回帰式

- 重回帰式では単回帰式に複数の独立変数を追加する。

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

変動する部分 定数

(式8.7) アパートの家賃 = $(b_1 \times \text{駅からの距離}) + (b_2 \times \text{築年数}) + (b_3 \times \text{部屋の広さ}) + b_0$

b : 各独立変数の偏回帰係数

他の独立変数を除外した、特定の独立変数の従属変数への影響力

→独立変数の単位が異なる場合は単純な大きさの比較はできない

※標準化偏回帰係数にすることで比較することは可能（7-2-3の偏相関と類似）

偏回帰係数の解釈 (留意点)

- ① 重回帰式において定数が大きい場合には、偏回帰係数が大きくても独立変数が従属変数に与える影響は小さくなってしまう。

(例) アパートの基準家賃が30万円の場合のほうが、3万円の場合と比べて偏回帰係数による家賃の変動の割合は小さくなる。

- 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、他の独立変数の予測力に影響を受ける場合がある。

→ 「各標準偏回帰係数の値 = 影響力の大きさ」だと見なすことは危険

重相関係数と決定係数

【重回帰分析における係数】

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 (coefficient of determination: R^2) :

重相関係数を2乗した値であり、独立変数全体でどの程度従属変数を説明しているかを示す。

重相関係数と決定係数

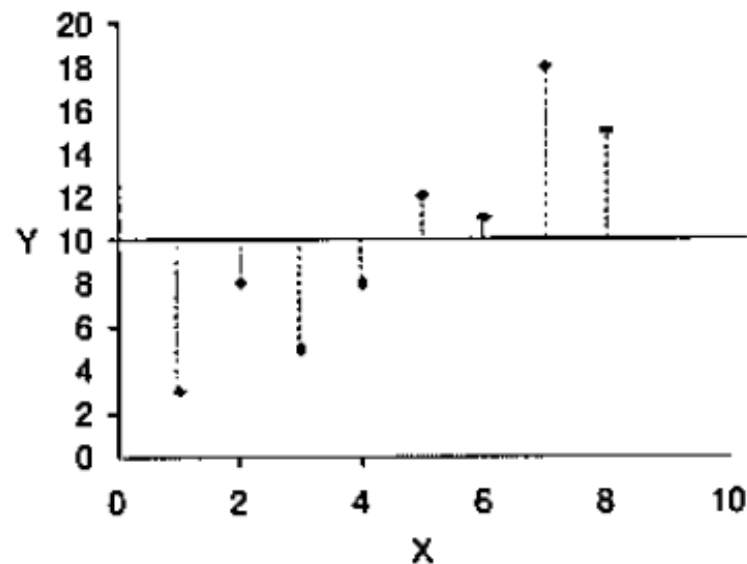
- 決定係数の算出では、3種類の分散が算出される。

① SS_T (total sum of square) :

個々の観測値と平均の差の2乗を足した全平方和。

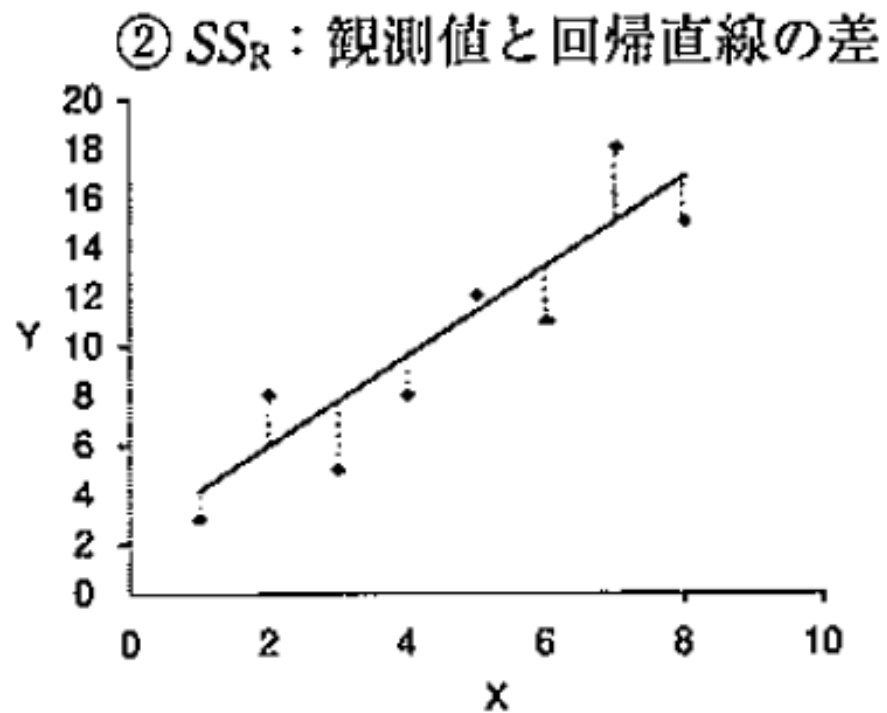
従属変数の平均が観測値のモデルとしてどの程度適切かを表す。

① SS_T : 観測値と平均 Y との差



重相関係数と決定係数

- ② SS_R (residual sum of square) : 残差平方和。
平均が回帰直線からどの程度ずれているか表す。

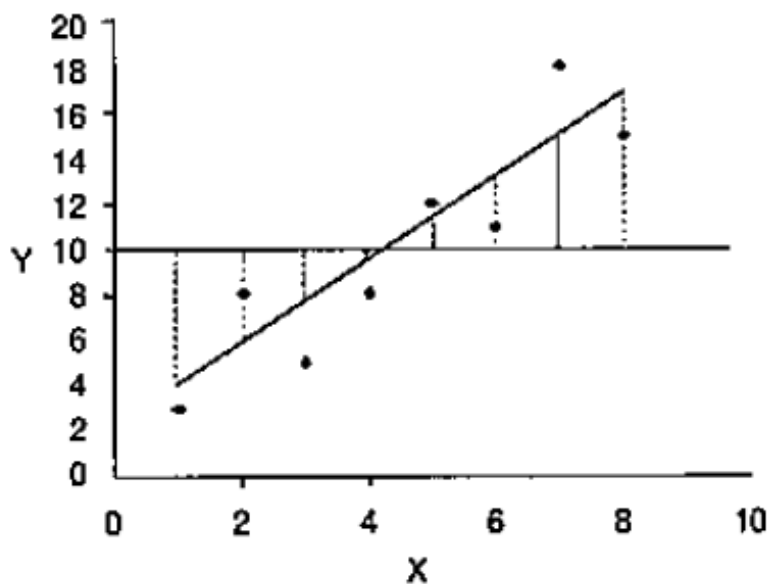


重相関係数と決定係数

③ SS_M (model sum of square) : SSTからSSRを引いた平方和。

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向上したかを示す。

$$\textcircled{3} \quad SS_M = (SS_T - SS_R)$$



重相関係数と決定係数

- 決定係数 (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

$$(式8.8) \quad R^2 = \frac{SS_M}{SS_T}$$

F検定 (分散分析)

- 決定係数が有効かどうかを分散分析で検定した結果は以下の図のように表示される。

分散分析^a

モデル		平方和	自由度	平均平方	F 値	有意確率
1	回帰	12542.863	1	12542.863	56.175	.000 ^b
	残差	6251.937	28	223.283		
	合計	18794.800	29			

a. 従属変数 期末テスト
b. 予測値: (定数)、勉強量。

Diagram labels: SS_R (Residual Sum of Squares), MS_M (Model Mean Square), SS_M (Model Sum of Squares), MS_R (Residual Mean Square), SS_T (Total Sum of Squares).

図 8.3 分散分析

- モデルの分散を自由度で割った平均平方和が、モデルの誤差分散を自由度で割った平均誤差分散よりどの程度大きいかを分散比 (F値) として算出し、その有意性を検定する。

$$(式 8.9) \quad F = \frac{MS_M}{MS_R} = \frac{12542.863}{223.283} = 56.175$$

回帰分析を行う際の注意点 (回帰分析の前提)

① サンプルサイズと質

厳密な基準はないが、回帰分析では比較的多くのサンプルが必要

【必要なサンプル数】

Tabachnick & Fidell (2007)

信頼性のある決定係数 : $50 + 8k$ (k = 独立変数の数)

各独立変数の有意性検定 : $104 + k$ (k = 独立変数の数)

回帰分析を行う際の注意点 (回帰分析の前提)

② 多重共線性

多重共線性 (multicollinearity) :

独立変数間に非常に高い相関がある場合に生じる問題

→本来関係ないはずの独立変数が従属変数の予測に貢献するなど

- 独立変数間の相関が.80以上なら、以下の指標を用いて多重共線性を診断する必要がある。
 - (1) 許容度 (tolerance) : 特定の独立変数を従属変数とし、他の独立変数群から予測した場合の決定係数を1から引くことで算出。.10以下だと多重共線性が発生。
 - (2) VIF (variance inflation factor) : 許容度の逆数。10以上だと多重共線性が発生。

回帰分析を行う際の注意点 (回帰分析の前提)

③ 外れ値

回帰直線は外れ値に大きく影響されるため、事前に調べる必要がある

【外れ値を調べる方法】

- (1) **残差** : 各データの残差を標準値に変換し、標準偏差 $\pm 2SD, 3SD$ 以上の値の割合を調べる。割合が全体の5%以内であれば問題なし。
- (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。1以上であれば問題あり。
- (3) **てこ比** : 各ケースにおける複数の変数データが平均からどの程度ずれているかを示す。値が大きい場合は問題あり。
- (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。値が大きいと外れ値である可能性あり。

回帰分析を行う際の注意点 – 回帰分析の前提 –

④ 残差の独立性、正規性、等分散性、線形性

回帰分析では、残差に関して以下の4つが満たされているという前提がある。

- (1) **残差の独立性** : どの独立変数の残差間にも差がないという前提。
ダービン・ワトソン法で調べることができる。
- (2) **残差の正規性** : 残差の散布図やヒストグラムを用いて調査。これが満たされない場合は非線形の回帰分析などに切り替える必要がある。

回帰分析を行う際の注意点 – 回帰分析の前提 –

- (3) **残差の等分散性**：独立変数に関わらず残差分散は等質性を有している必要がある (回帰直線に沿って散らばっていることが望ましい)。
- (4) **残差の線形性**：線形回帰分析の場合、残差は予測値と線形関係にある必要がある。標準化残差と標準化予測値の関係を散布図にすることで調べられる。

投入法

複数の独立変数から求められた重回帰式の有意性→**F検定**

個々の独立変数が従属変数の予測に寄与するか→**t検定**

- 重回帰分析では、従属変数の予測に用い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及び偏回帰係数が変化する。



◎ 目的に合った投入法を用いることが重要

投入法

① **強制投入法**：全ての独立変数を同時に投入して従属変数の予測を行う。

※関係の無い独立変数を投入した場合も決定係数は大きくなるため、理論や仮説に基づいて慎重に独立変数を選択する必要がある。

② **階層的投入法** (階層的回帰分析)：理論や仮説に基づいて1つずつ独立変数を投入する方法。理論的に優先する独立変数の説明率を調べるために用いる。

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。

※あくまで統計的根拠に基づいて投入されるため、理論に適っているかは別途判断する必要がある。

投入法 – その他の投入法 –

変数増加法：ステップワイズ法同様に独立変数を順に投入していく方法。
投入するごとに除去すべき変数がないかを分析することはできない。

変数減少法：最初にすべ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。



ステップワイズ法は第2種の過誤を生じる可能性があるため、変数減少法の方が望ましいという意見もある (Field, 2009)

実践

- ▶ 8-3-1, 8-3-2 強制投入法
- ▶ 8-3-3, 8-3-4 ステップワイズ法
- ▶ 8-4 ダミー変数を使った回帰分析