

SPSSによる重回帰分析

村瀬 洋一

1. 重回帰分析とは何か

1.1. 目的と具体例

◆重回帰分析の目的 — 説明変数Xを複数設定し、被説明変数Yとの関連が強いのが、どの変数なのかを解明すること。

相関係数や2重クロス集計のように、表面的な2変数の関連を見るだけでなく、他の変数の影響を取り除いた後（統制後、コントロール後）の関連を解明するのが目的。

線型の関連（回帰直線）を前提に、関連の強さを測る。

説明変数、被説明変数とも連続変数（量的変数）の場合に用いる。説明変数が離散変数（カテゴリー変数）の場合は、分散分析を用いる。なお、両方とも離散変数の場合は、クロス集計やログリニア分析を用いることになる。

なお、Yが2段階回答の場合は、ロジスティック回帰分析（あるいは判別分析、数量化2類など）という別の分析法を用いる。

社会調査の4段階回答などは、厳密には連続変数ではないが、3段階以上ならば、連続変数とみなして用いることが多い。データ人数が多い場合（おおむね400人以上など）は、このような考え方で、とくに問題はない。

◆具体例

例えば調査で「生活全般満足度」に関する問の、4段階回答があった場合に、この原因を解明することを重回帰分析の目的とする。この変数をYとして、この原因となっている変数は何かを複数考え、複数のXを用いて重回帰を行うことになる。

社会調査の場合、Xとして年齢、教育年数、財産や収入などの属性変数や、自営業かどうかなどの01のダミー変数、その他の心理的変数（態度や意識）を用いることが多い。

もし、他の変数の影響を取り除いても（コントロール後でも）、年齢がYと関連していた場合、年齢がYの原因となっている、と考えることができる。実際、高齢の人ほど伝統的価値観を持っているため、そのような結果になることは多い。しかし、高齢の人は低学歴な傾向があるため、表面的にはYと学歴も関連があるように見える。そのような表面的関連でなく、他の変数の影響を取り除いた後の、真の関連を見つけることが分析の目的である。

分析結果として、まず偏回帰係数をいくつか出し、どれが大きいかを確認するとよい。

1.2. 重回帰分析の考え方 (ポーンシュテット・ノーキ, 1990, 第8、第11章などを参照)

説明変数が2つの場合の重回帰分析のモデル

$$\text{標本回帰式 } Y_i = a + b_1 X_{1i} + b_2 X_{2i} + e_i \quad \dots \dots \dots (1)$$

$$\text{標本予測式 } \hat{Y}_i = a + b_1 X_{1i} + b_2 X_{2i} \quad \dots \dots \dots (2)$$

Y_i : i番目の個体の、被説明変数Yの値

\hat{Y}_i : i番目の個体の、被説明変数Yの予測値

X_1 、 X_2 : i番目の個体の、説明変数 X_1 、 X_2 の値

a : 切片

b : 偏回帰係数

e_i : 誤差項 (残差項)

$$e_i = Y_i - \hat{Y}_i \quad \leftarrow \text{重要!}$$

bは回帰直線の傾きの大きさを表している (村瀬他, 2007:123の図を参照)。

この数式は、以下にある図1のようなモデルを表しているにすぎない。

1.3. 回帰分析のパラメーターの推定と解釈

a 、 b_1 、 b_2 の値の推定 → 最小自乗法 (Ordinary least squares OLS) を用いる。

$\sum e_i^2$ を (誤差の二乗の合計を) 最小にするように、 a 、 b_1 、 b_2 を推定する。

・ 偏回帰係数 (b_1 、 b_2)

他の変数の効果を統制した上で (統計的コントロールの後で)、説明変数が1単位変化した場合、被説明変数がどのくらい変化するかを示す。

・ 標準偏回帰係数 (β_1 、 β_2 ベータ係数、ベータ加重)

XとYを標準化した (Z得点にした) 上で求めた回帰係数

説明変数が1標準偏差増えた場合、被説明変数がどのくらい変化するかを示す。

ただし、上記の係数は、その説明変数に固有の値ではない。他の説明変数が変われば、当該の説明変数の係数値も変わる。(久米・飯塚、1987、p.99-を参照)

回帰係数は相関係数とは異なる。他の変数Xの影響を除いた場合の、Yとの関連の強さを表しているのである。

1.4. 決定係数 (R^2)

説明変数Xが、被説明変数Yの分散をどのくらい説明しているかを示す。モデル全体 (回帰式全体) の説明力を表す。レンジは0~1。モデルで分散を完全に説明しているときは1になる。

$$\begin{aligned} R^2 &= \frac{\sum [(Y_i - \bar{Y})^2 - (Y_i - \hat{Y}_i)^2]}{\sum (Y_i - \bar{Y})^2} = \frac{SS_{TOTAL} - SS_{ERROR}}{SS_{TOTAL}} \\ &= \frac{SS_{REGRESSION}}{SS_{TOTAL}} \\ &= \frac{\text{モデルで説明できる分散}}{\text{全分散}} \quad (3) \end{aligned}$$

$$SS_{TOTAL} = SS_{REGRESSION} + SS_{ERROR} \quad \dots \dots \dots (4)$$

(平均値と観測値の距離) = (回帰モデルで説明できる距離) + (観測値と予測値のずれ)

数式の各項が、村瀬他(2007:125)の図ではどの部分になるか、理解できるとよい。

なお、説明変数が2つの場合、標準偏回帰係数と決定係数の間には以下の関係が成り立つ。

$$R^2 = \beta_1^* r_{YX1} + \beta_2^* r_{YX2} \quad \dots \dots \dots (5)$$

r_{YX1} : Yと X_1 の相関

r_{YX2} : Yと X_2 の相関

1.5. 決定係数の有意性検定

R^2 の有意性検定は、F検定によって行う。

$$F = \frac{MS_{REGRESSION}}{MS_{ERROR}} = \frac{\text{モデルで説明できる分散}}{\text{モデルで説明できない分散}} \quad \dots \dots \dots (6)$$

$$MS_{REGRESSION} = \frac{SS_{REGRESSION}}{\text{自由度}}$$

自由度 : 説明変数の数

$$MS_{ERROR} = \frac{SS_{ERROR}}{\text{自由度}}$$

自由度 : $N - 1 - (\text{説明変数の数})$

1.6. 多重共線性(マルチコ)

重回帰分析はとても分かりやすく有効な分析法だが、説明変数X同士の相関が高い場合は、重回帰分析を行うことはできない。この点によく気をつけること。

説明変数間の相関がとても高い場合、回帰モデルは非常に不安定になる。これは、説明変数の間にすでに別の線形回帰関係が含まれているということであり、その意味でこのような現象を「多重共線性(multicollinearity)」と言う。経験的に、説明変数間の相関が0.7以上ならば危険であると言われている。

多重共線性に注意するために、回帰分析を行う際には、まず説明変数間の相関行列を見て、相関がとても強いものがあれば、片方は説明変数から除く、といったことが必要である。どのような説明変数の組み合わせがもっとも適しているかを明らかにするために、変数選択の方法がいくつか考えられている。詳しくは、村瀬他(2007)などを参照。

1.7. 回帰分析を行う上での注意(久米・飯塚, 1987:193-)などを参照)

説明変数が多ければ多いほど、決定係数 R^2 は必ず大きくなる。しかし、決定係数が大きい回帰式が良いモデルというわけでは、まったくない。

上記のように、説明変数の間に相関が強いと、回帰分析はできない。極端なことをいえば、説明変数が2つで、その間の相関が1ならば、2つの説明変数は同じものなのだから、どちらか1つを回帰モデルに入れば良いのである。モデルはシンプルほど良い。

決定係数 R^2 がどれくらい大きければ良いと見なすのには難しい問題だが、社会調査データでは、測定誤差などのノイズが多いので、0.20を越えればとても大きいと言うことが多く、値が0.10以下でも、モデル全体のF値が有意ならば（統計的に意味があるならば）良い。心理学や経済学で用いるデータでは0.50以上になることが多いが、社会調査データでは、そのような例は少ない。

なお、重回帰分析を行うためには、ある程度のデータ数が必要である。通常、400人以下のデータで重回帰分析を行うことはない。最低でもおおまかにいえば、少なくとも200人は必要である。400人以下の時は、自由度調整済み決定係数を用いた方がよい。

◆問 重回帰分析の説明変数Xは連続変数だが、離散変数も説明変数として用いることができる。どのようにすれば良いか。

答 if文を使って、0、1型の変数（ダミー変数）を作ればよい。

2 カテゴリーの時 Q1 (0, 1) 例 男なら1、女なら0

3 カテゴリーの時 QW (0, 1) 例 ホワイトカラー、ブルーカラー、農業
QB (0, 1) 3 カテゴリーの場合、ダミー変数を2つ作ればよい。

3 カテゴリーの時に、ダミー変数を3つ作らないように注意。

1.8. ダミー変数を作るSPSSシンタックスの例

COMPUTE文やIF文を使って、ダミー変数を作ることができる。

書式 IF (条件式) 条件があった場合の命令文 ピリオド

例 職業の変数から、自営かどうかの0、1型変数を作成

COMPUTE JIEI=0. ←新変数JIEIを作成（変数名は好きな名前でもよい）
IF (Q19BS2=6) JIEI =1.

2. SPSSによる分析

2.1. SPSSの操作

シンタックスを書くか、もしくはSPSSで、画面上の「分析」をクリックし、「回帰」→「線形」を選ぶ。

シンタックス例

```
REGRESSION                                ← (REGだけでも動く)
  /DEPENDENT Q6B
  /METHOD=ENTER AGE Q2 Q3 Q4A .
```

解説

```
/DEPENDENT      この行にYになる変数名を書く
/METHOD=ENTER この行にXとなる変数名を複数書く
/MISSING LISTWISE 欠損値がある人のデータは取り除く（この行は省略可能）
```

分析実行前に必ず欠損値処理をすること。年齢や教育年数、自営ダミーなど、基本属性もXとして入れるとよい。変数は逆転するなど方向を事前にそろえること。

2.2. 作業手順

まず、被説明変数Yを1つ決め、さまざまなXを入れて自分の好きなモデルを考える。

はじめは数個の説明変数Xを入れ、少しずつ増やしてみると良い。ただし、最終的には、すべてのXをいれたモデルを検討すること。一部のXだけを入れた分析結果を、いろいろ出して表にしても、とくに意味はない。

分析の前に必ず欠損値処理を行う。また、変数の方向をそろえる（回答方向を逆転した新変数を作るなどする）。YもXも量的変数しか使えないことに注意。

良いモデルを得るために多重共線性に注意せよ。まず、事前に説明変数間の相関行列を見てみることに。

性別ダミー変数を使うか、あるいは男女別にデータを分割してから分析し、2つの結果を比べるなどとよい。調査データの場合、分割した方がうまくいくことが多い。

2.3. 結果のまとめと解釈

分析結果は、学術論文では以下のような形式の表にまとめる。図の方が一般向けには分かりやすい。各説明変数の偏回帰係数は有意か、モデル全体の説明力はどうか、なぜそのような結果が出たのかなどについて検討し、結果の解釈や考察を行うこと。

この例では3つのモデルについて表している。Yとの相関係数rは別途分析すること。

重回帰分析の結果 表のまとめ方の見本
(数字は架空例)

説明変数 () 内は変数のレンジ	地方議会議員		町内会役員		企業の経営者	
	β	r	β	r	β	r
年齢(20-69)	0.03	0.14**	0.03	0.11**	-0.01	-0.02
学歴(教育年数 6-17)	-0.01	0.00	-0.10	-0.07**	0.05	0.03
世帯資産(保有財産数 0-20)	0.15	0.14**	0.12	0.12**	0.14	0.14**
居住地域都市度(1-8)	-0.18	-0.12**	-0.08	-0.06**	-0.01	0.00
地域移動経験の有無(1,0)	-0.42**	-0.06**	-0.35**	-0.05**	0.16	0.00
組織内の役職(1-6)	0.13	0.05**	0.07	0.01	0.34**	0.12**
従業先企業規模(1-7)	-0.01	0.00	0.07	0.03*	-0.24**	-0.13**
本人職業威信スコア(26.7-83.5)	0.01	0.02	0.00	0.00	0.01	0.00
父職業威信スコア(23.4-87.3)	-0.02	0.00	-0.04	-0.02	-0.03	0.00
父学歴(教育年数 6-17)	0.00	0.00	0.00	0.00	0.00	0.00
本人職 自営ノンマニュアル(1,0)	0.68**	0.05**	0.72**	0.05**	0.15*	0.00
本人職 自営マニュアル(1,0)	0.29	0.00	0.36*	0.02	-0.08	0.00
本人職 農業(1,0)	0.42*	0.03*	0.85**	0.07**	-0.79**	-0.13**
R-square	0.27**		0.21*		0.16*	
Adjusted R-square	0.21		0.18		0.12	
N	381		324		356	

注 被説明変数は、××の場合4、××の場合0
縦1列が1つの回帰式を表し、点線は標準偏回帰係数と相関係数
説明変数のうち、レンジが(1,0)のものはダミー変数。職業ダミー変数の基準はその他の職業
** 1%水準で有意 * 5%水準で有意

注意点

表だけを見て、第3者が分かるのが大原則である。

表タイトルも的確に分かりやすく。表タイトルは表の上書き、表番号をつける。

説明変数についての説明を、表の下に注で書く。

通常、縦1列が1本の重回帰式になる。この例では3本の重回帰分析の結果を1つの表にまとめている。

説明変数間の相関行列も、別途表にすると良い。

有効桁は2桁でよい。あまり細かい数字を書いても意味がない。

★SPSSの出力をエクスポートしてテキスト形式で保存し、エクセルで読み込んで有効桁などを合わせ表にすると簡単。あるいは、エクセルでテキストファイルを開いてから画面上の「データ」をクリックし、区切り位置を指定する。

四捨五入したい範囲のセルをマウスで囲んで「書式」をクリックし、「セルの書式設定」→「ユーザー定義」を選択すると、数字を小数点以下2桁等に揃えることができる。

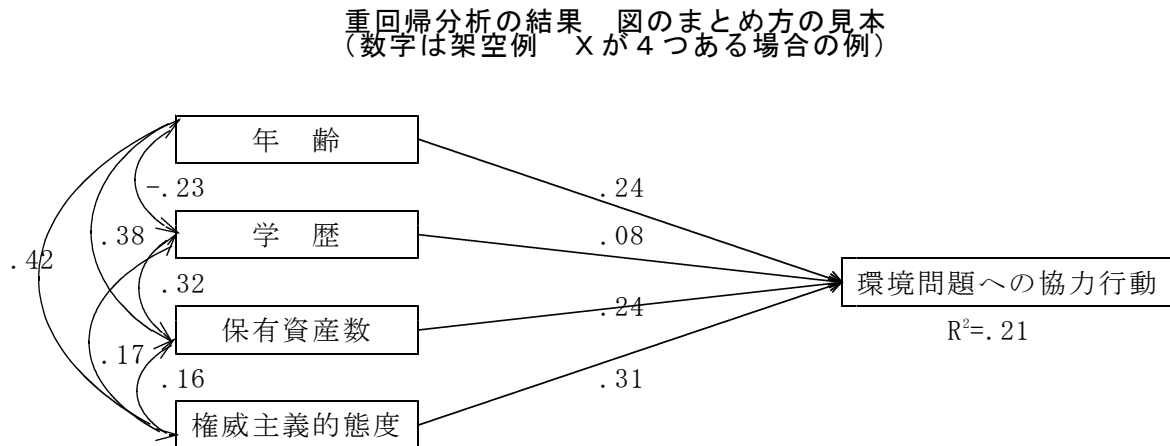


図1. 向環境行動の規定因に関する重回帰分析結果 1998年××調査
数字は標準化係数

図を書く際の注意点

因果関係の流れは、左から右へ、原因から結果となるようにする。

実在する変数（実在変数、観測変数）は四角で表現（因子は楕円で表現）。

決定係数 R^2 も必ず書く。

図のタイトルは図下を書く。数字の説明も忘れずに書く。

説明変数間の相関も書くこと（相関係数の分析で出力）。

なおワードでは、画面上の「挿入」をクリックし、図形ボタンが表示されるので、ボタンをおして矢印などを書くといい。詳しくは以下を参照。

2.4. SPSS出力の見方について

モデル全体の決定係数はR二乗の値を、各変数の標準偏回帰係数（standardized estimate）はベータを見れば良い。各値の有意水準（有意確率、危険率）も見ること。

R二乗の有意水準（モデル全体の有意水準）は、F値の有意水準を見ればよい。これが0.05未満ならば、R二乗が誤差である確率は5%未満なので、このモデルを採用して良い。

2.5. 変数選択

重回帰分析に説明変数を複数入れ、その後、どの変数を採用するのが適切かを検討することができる。このことを変数選択という。初めは、10個の説明変数でモデルを作り、その後、説明変数を5個くらいに絞るなどするとよい。

調査データの場合、とくに変数選択をせず、強制投入法としてすべてのXを用いることも

多い。

- STEPWISE

既存のモデルをもとに、次に新しい変数を入れるか、あるいはモデルに既に入っている変数を落とすかを逐次的に行う。

- RSQUARE

候補となる説明変数のすべての組み合わせについて、回帰式と変数選択のために提案されている各種統計量を計算する。

3. 分析時の注意点

3.1. 分析の前に必ず欠損値処理をすること

多くの場合、欠損値は9か99。SPSSの場合、missing valuesコマンドを用いる。

回答が2桁の場合、欠損値99である。まず単純集計をとって確認するとよい。

3.2. 分析の前に変数の向きを必要に応じて逆転し、わかりやすく設定する

分析を行う前に、原則として、すべての変数を、数字が大きいほど肯定になるように直すこと。数字が小さいほど肯定となる変数が混ざっていると、とても分かりにくい。

シンタックスのデータ定義文の後で、以下のようなCOMPUTE文を書けばよい。

例 そう思う1 --- そう思わない4 → そう思う4 --- そう思わない1

Q4a を逆転し、新変数N4a (好きな名前をつける) を作る

```
COMPUTE N4a = 5-Q4a .
```

3.3. 用いる変数について

重回帰分析で用いる変数は、XもYもすべて連続変数(量的変数)であることに注意。名義尺度の変数はダミー変数以外は使えない。また、変数内でカテゴリー合併などをする必要はない。むしろ、なるべく回答の段階は細かい方が、連続変数に近くなるのでよい。

4段階尺度や順序など、厳密には連続変数ではないが、量的変数と見なして重回帰で使って問題はない。ただし被説明変数Yは、3段階以上が望ましい。

カテゴリー変数(質的変数、離散変数)をXとしたい時は、if文やrecode文を用いて、ダミー変数や連続変数に直すとよい。

例 問16の学歴を、教育年数という連続変数に直す。その他7、無回答9は欠損値

```
COMPUTE      EDU=Q16.
```

```
RECODE      EDU(1=6)(2=9)(3=12)(4=13)(5=14)(6=16)(7,9=99).
```

```
MISSING VALUES EDU(99).
```

4. 発展版

4.1. 男女別等の分析 - ファイルの分割について

調査データの場合、男女別に分析して結果を出すことが多い。重回帰分析も、多くの場合、男女別にデータを分割した後で分析すると、より明確な結果が出る。男女すべて合わせたデータだと、うまく関連が出ないこともある。

SPSSにはデータ分割機能があるので便利である。SPSSのデータウィンドウで、画面上の「データ」をクリックし、ファイルの分割を選ぶと便利。データを男女別に2つに分割してから、分析を行うと、男女について2つの分析結果が出る。

4.2. モデルの作り方

自分の目的を明確に決めてYとなる変数を1つ設定する。Xとして、心理的変数（意識に関する質問項目など）と社会的変数（年齢、学歴、財産数など基本属性や社会的地位に関連するもの）を入れ、さまざまなモデルを作ってみると良いだろう。

初めは、年齢や学歴など基礎的なXだけを入れたモデルを作り、少しずつXを増やしていくとよい。因果関係をよく考えてXを入れると良い。

最終的なモデルは、Xとして心理的変数と社会的変数の両方を含むと良い。重回帰分析は、多くのXを同時に投入することに意味がある。別々に入れたモデルはあまり意味がない。ただし多重共線性には注意する。

4.3. エクセルやワードでの図表の書き方

- ・ワードやエクセル画面の上「挿入」をクリック
- ・図形描画ボタン、またはオートシェイプを選ぶ
- ・ボタンを押して四角や線や矢印などをかく。

1) 図を微調整したい時

かいた図を右クリックして書式設定を選ぶ。線の太さや矢印種類などを変更できる。両方向矢印などにすることができる。

2) 図の中に文字を書くには

- ・図形を右クリック → テキストの追加
- ・図形を右クリック → オートシェイプの書式設定
「色」ボックスをクリック → 塗りつぶしなし（白でなく透明になる）
- ・あるいは画面上「挿入」をクリックすると「テキストボックス」が出る。設定して文字を書く。その後、微調整は、テキストボックスを右クリックして書式設定を選ぶ

3) エクセルでの罫線の引き方

- ・線を引きたいセルをマウス（またはシフトキー＋矢印）で囲む
- ・囲んだ部分を右クリックして「書式設定」→「罫線」タブを選ぶ
- ・下線ボタンなどを押しOKボタンを押す

4) 曲線矢印の引き方

- ・画面上「挿入」を押し、図形ボタン押し曲線を選ぶ
- ・曲線を引く。真ん中で一度クリックしさらに引く。書き終わるときはダブルクリック
- ・引いた線を右クリックし「書式設定」を選ぶ
- ・矢印「始点や終点のスタイル」を選ぶ

4.4. エクセルで作った表をワードの中に貼るには

- ・まずエクセルで表を作る
- ・ワード画面で画面上の「挿入」をクリック
- ・「オブジェクト」を選択し「エクセルワークシート」新規を選ぶ

- ・ワークシートが出てくるので、自分で作った表をはりつける

新規でなく「ファイルから」を選び「参照」ボタンを押して既存ファイルを読み込んで良い

4.5. 論文の構成 ー 一般的な分析の流れについて

レポートや論文を作る際には、冒頭で目的（何を明らかにしたいか）を明確に書く。そしてYとなる変数を1つ決める。その後、まず因子分析結果や相関行列を出し、全体的な変数間の関連を確認すると良い。その後、因果関係を自分の頭で考えて、何をXにするかを決めてモデルをいくつか作り重回帰分析で因果関係を確認する。その後、さらに用いる変数を絞って、クロス集計やエラボレイションを行うと良い。

論文には、分析結果として、基本的な男女別集計の横棒グラフ等をまず載せ（分布の偏りを確認し、どのような質問項目か読者に分かってもらう）、相関行列（または因子分析）、重回帰分析、主要な変数に関するクロス集計の順で結果を並べることが多い。

5. 課題

自分で自由にテーマを決め、何らかの調査データを用いて、男女別に重回帰分析を行う。結果を、男女別の2つの図にまとめ、自分の解釈を書く。被説明変数Yは、自分が興味ある質問項目を1つ決めればよい。説明変数Xを5個以上入れること。

上記「分析時の注意点」に、十分に気をつけること。

6. 参考文献

ボーンシュエット・ノーキ. 1990. 『社会統計学 ー 社会調査のためのデータ分析入門』. ハーベスト社.

早川毅. 1990. 『回帰分析の基礎』朝倉書店.

市川伸一・大橋靖雄. 1987. 『SASによるデータ解析入門』. 東京大学出版会.

石村貞夫. 2001. 『SPSSによる多変量データ解析の手順』東京図書.

石村貞夫. 2001. 『SPSSによる統計処理の手順』東京図書.

久米均・飯塚悦功. 1987. 『回帰分析』. 岩波書店.

蓑谷千風彦. 1990. 『回帰分析のはなし』東京図書.

縄田和満. 1998. 『Excelによる回帰分析入門』朝倉書店.

三宅一郎・山本嘉一郎他. 1986. 『新版SPSS X 基礎編』東洋経済新報社.

村瀬洋一他編. 2007. 『SPSSによる多変量解析』オーム社.

室淳子・石村貞夫. 2002. 『SPSSでやさしく学ぶ多変量解析』東京図書.

岡太彬訓・古谷野亘. 1993. 「多変量解析法の不適切な利用」. 数理社会学会
『理論と方法』Vol. 8 No. 2.

小塩真司. 2004. 『SPSSとAmosによる心理・調査データ解析 ー 因子分析・共分散構造分析まで』東京図書.

佐和隆光. 1990. 『回帰分析』朝倉書店.

田部井明美. 2001. 『SPSS完全活用法 ー 共分散構造分析(Amos)によるアンケート処理』東京図書.

★おまけ シンタックス例

```
***** ダミー変数作成 *****/
```

```
COMPUTE      MOTIIE =0.
```

```
IF (Q25 =1)  MOTIIE =1.
```

```
***** 変数の方向を逆転 *****/
```

```
MISSING VALUES  Q7A (9).
```

```
COMPUTE      N7A =5-Q7A.
```

```
***** 重回帰分析 *****/
```

```
REG
```

```
  /DEPENDENT Q29
```

```
  /METHOD=ENTER AGE EDU Q3 Q4 Q5.
```

```
***** 分散分析 *****/
```

```
UNIANOVA Q16 BY Q27 NENDAI
```

```
  /METHOD=SSTYPE(3)
```

```
  /PLOT=PROFILE(NENDAI*Q27)
```

```
  /PRINT=DESCRIPTIVE
```

```
  /DESIGN=Q27 NENDAI Q27*NENDAI.
```

```
***** 欠損値を除く処理 *****/
```

```
SELECT IF age < 99.
```

★モデルの作り方

Xは何個くらいがいいのか

とくに基準はないが、社会学的データの場合、普通は年齢、学歴、収入と基礎的な社会意識項目数個の他に、関連する項目を数個いれ、10個前後のXを入れて分析することが多い。

「伝統的価値観」に関する項目など、何らかの基礎的態度と関連する項目を、統制変数として入れた方がよい

有意でないXを除いた方がいいのか

とくに除く必要はない。関連がないということも、重要な発見

Yと有意な関連があるXはいくつくらいあった方がよいのか。

1つでも構わない。すべてがYと無関連だと、他にXを探すべきということになるが、基本的に、無関連でも構わない。

ステップワイズの使い方は

社会学的データの場合、とくに使う必要はない。ただ、X同士の相関が強い場合は、どちらを投入すべきか判断するために、部分的に使うことがある。