「回帰分析（1）：考え方」

**■クラスター分析の小課題について**

クラスター分析の手続き自体はほとんどの人ができていた。

ただし、クラスターの数をどうやって決めたのか書かれていない人が多くいた。

そのほかの点で、わりと共通していた問題点は以下のとおり。

・図表に番号とタイトルがない。

必ず「図 1　各クラスターの平均得点による折れ線グラフ」といった図表番号とタイトルを付けて、本文中では「……についてまとめると図 1 のようになる。……。図 1 のクラスター1 を読み取ると……」というように、図表番号で参照する。

・分析の結果と考察が区別されていない。

客観情報（結果）と、主観的解釈（考察）を区別することは非常に大切。

・最後に全体の「まとめ」がない。

「まとめ」では、分析結果だけをまとめるのではなく、何をしようとしてという所（目的や方法）から振り返って、簡潔にポイントをまとめる

例）今回の報告では、大学生の幸福感の軌跡を明らかにするために、各年齢での幸福感を得点化してもらう調査を行なった。45 名の大学生のデータをクラスター分析した結果、○○ということがわかった。この結果について、××の視点から考察し、△△という結論に至った。

**■回帰分析の目的と魅力**

今回からは、**回帰分析**（regression analysis）について解説する。回帰分析は、ある 1 つの変数（従属変数）の値を、他の変数（独立変数）の値で説明しようとするときに、もっとも頻繁に利用される分析技法である。たとえば、ある大学の先生が学生の遅刻に頭を悩ませているとする。遅刻の原因として、アルバイトのやりすぎによる疲れがあるのではないかと考え、15 人の学生に調査をしたとする。1 ヶ月のアルバイト時間を X 軸に、遅刻回数を Y 軸にして図 1 のような散布図を描くと、その関係性がわかる。このとき、散布図の上に直線を引いてみたくなることがある。

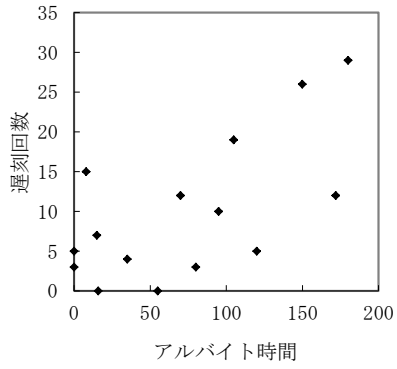



図 1　アルバイト時間と遅刻回数の関係（仮想データ）

このような直線を引きたくなるのは、次のように考えているからである。「X と Y の関係は、本来この直線のような関数で表せるのではないだろうか。実際のデータがこの直線からいくらかずれているのは、何らかの誤差によるものだろう」と。より定式的に書けば、「本来の Y の値は、X の値から $\hat{Y}=\alpha+\beta X$ という直線の関数で表せる（$\hat{Y}$ は実際の Y の値ではなく、予測値としての Y の値を表す)」と考えていることになる。α は直線と Y 軸が交わる切片を表し、β は直線の傾きを表す。α や β は定数なので、具体的には $\hat{Y}=4.5+0.1X$ といった形で Y の予測式は表される。上のような予測式のことを**回帰式**（regression equation）と呼び、回帰式によって表される線のことを**回帰線**（regression line）と呼ぶ。また、回帰式の α を定数項、β を**回帰係数**（regression coefficient）と呼ぶ。

回帰分析の目的は、回帰線を最適に調整することを通して、ある変数（従属変数）の値が、その原因と考えられる変数（独立変数）によってどのように説明できるのかを統計的に明らかにすることである。何らかの因果関係を想定して、その関係性の有無や方向、強さに関心を持つことは極めて一般的な問題意識であり、その疑問に正面から答えを出してくれることが回帰分析の魅力である。

回帰分析の重要な手続きは、次の 3 点にまとめられる。順に説明しよう。

（1）もっともよい線を引く。
（2）その線はどのくらいよい線であるかを評価する。
（3）母集団についても同様の線を引く価値があるかどうかを判断する。

**■概要をスライドで確認**

・テキスト p. 99 の図が回帰分析の本質。
・実際のデータで最適な回帰式を求めると、p. 100 のようになる。
・分析の結果を図に戻すと……

**■最適な回帰式を推定する**

回帰分析の最初の手続きとして、散布図の中にもっともよい回帰線を引かなければならない。もっともよい回帰線とは、実際のデータと予測値との差、つまり $Y-\hat{Y}$（残差と呼ぶ）の合計がもっとも小さくなる線である。ただし、残差がプラス側かマイナス側であるかは、問題ではないので、残差を 2 乗した値を用いて、その合計値が最も小さくなるようにする。この合計値を**残差平方和**（residual sum of squares）と呼ぶ。残差平方和が小さいほど、その回帰線はよい回帰線と考える。

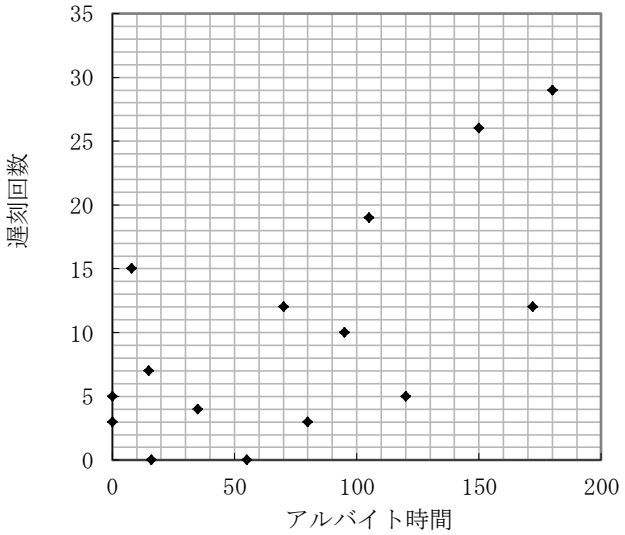
（練習）**※次ページをみないように！**

1．自分が最適だと思う直線を散布図の上に引いてみよう。

2．その直線の切片と傾きを読み取って、式に表わしてみよう。

$\hat{Y} = \alpha + \beta X$

→ $\hat{Y} =$ _____ + _____ $X$




3．自分が引いた直線について、残差平方和を求め、周りの人と比較してみよう。（残差平方和が小さいほどよい回帰線ということになる）

| | アルバイト時間 $X$ | 遅刻回数（観測値） $Y$ | 自分が引いた直線 | | |
|---|---|---|---|---|---|
| | | | 予測値 $\hat{Y}$ | 残差 $Y-\hat{Y}$ | 残差平方 $(Y-\hat{Y})^2$ |
| 1 人目 | 55 | 0 | | | |
| 2 人目 | 35 | 4 | | | |
| 3 人目 | 180 | 29 | | | |
| 4 人目 | 172 | 12 | | | |
| 5 人目 | 150 | 26 | | | |
| 6 人目 | 8 | 15 | | | |
| 7 人目 | 80 | 3 | | | |
| 8 人目 | 95 | 10 | | | |
| 9 人目 | 0 | 3 | | | |
| 10 人目 | 15 | 7 | | | |
| 11 人目 | 16 | 0 | | | |
| 12 人目 | 120 | 5 | | | |
| 13 人目 | 105 | 19 | | | |
| 14 人目 | 70 | 12 | | | |
| 15 人目 | 0 | 5 | | | |

（合計）↓

残差平方和 ＝ _____

目分量で適当に引いても、そこそこよい回帰線が引けると思われるが、数学的には微分方程式を解くことで最適な線を導くことができる。このように数学的に最適な回帰線を求めることを**最小二乗法**（method of least squares）と呼ぶ。

数学的な詳細は省略するが、方程式を解くと、具体的にいまのデータの場合には、$\beta$＝0.095、$\alpha$＝3.01 が最適である。つまり、$\hat{Y}=3.01+0.095X$ という回帰式最適である。このとき、残差平方和は 632.13 になり、他にどんな回帰線を考えても、これよりも小さな残差平方和をとることはない。

この回帰線から、次のように具体的な意味を読み取れる。アルバイトをしていない場合（X が 0 の場合）は遅刻の回数が 3.01 回と予測され、アルバイト時間が 1 時間増えるごとに、0.095 回ずつ遅刻の予測回数が増える。

**■回帰線の説明力を評価する**

最小二乗法によって、最適な回帰線は求まる。しかし、最適な回帰線であったとしても、従属変数の予測に十分な説明力（予測力）を持つとは限らない。もともと独立変数に従属変数を説明する力がない場合には、最善を尽くしても十分な説明ができるはずはないからである。そこで、2 つ目の手続きとして、その回帰線はどのくらいよい線であるか、説明力の強さを評価する。

回帰線が持つ説明力の評価は、一般に**決定係数**（coefficient of determination）によってなされる。決定係数は 0～1（0%～100%）の値を取り、独立変数で従属変数の値をどれだけ説明できるか、その割合を表す。

決定係数は次のような考え方に基づいている。いま、従属変数（Y）の予測のために独立変数（X）の情報を用いることができないとしよう。つまり、1 人 1 人のアルバイト時間が分からない中で、遅刻回数をなるべくずれが少ないように予測することを考える。このとき、最適な予測方法は、常に Y の平均値を予測値として用いることである（図 2)。

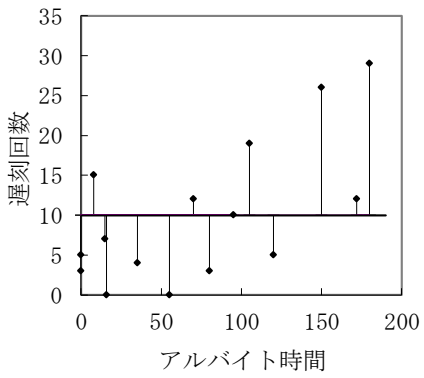



図 2　独立変数を利用しない最善の予測

このときの残差平方和を 100%として、最適な回帰線を用いると残差平方和が何%減少するかを算出したものが決定係数である。今回のデータの場合、遅刻回数をすべて平均値で予測すると、残差平方和が 1124.00 になる。最適な回帰線による残差平方和は 632.13 だったわけなので、この回帰線によって残差平方和は 491.87 だけ改善した。これは、もともとの残差平方和の 43.8%にあたる(491.87/1124.00=0.438)。つまり、決定係数 $R^2$＝0.438

で、遅刻回数の 43.8%がアルバイト時間によるこの回帰式で説明できることがわかる。

ただし、実際のデータ分析では、さらに調整を加えた**調整済み決定係数**（adjusted $R^2$）を用いることが多い。決定係数は、母集団における実際の説明力よりもわずかに大きくなる偏りを持つ。この偏りは、標本の回答者数が少ないときなどに、無視できないほど大きくなるので、決定係数をやや小さく調整し直すわけである。

今回の回帰分析の場合、決定係数は 0.438 だが、調整済み決定係数は 0.395 となる。結局、遅刻回数の 39.5%がアルバイト時間を原因と考えることで説明できることが分かる。決定係数と調整済み決定係数の値がやや大きくかけ離れているのは、標本の人数が 15 人と非常に少ないためである。通常の調査データでは、それほど大きな違いは現れない。

決定係数がどのくらい大きければ十分なのか、明確な基準はない。学問分野や分析対象、分析目的によって必要な説明力は異なるからである。一般的には、社会調査のデータ分析で求められる説明力（決定係数）の水準は、あまり高くないことが多い。10%を切っていても有意義な分析とみなされることも珍しくはない。

**■説明力を統計的に検定する**

最後に残された手続きは、この最適な回帰線で、母集団についても説明すべきかどうか判断することである。つまり、回帰線の説明力が統計的に有意かどうかを検定する。最適な線を求め、それがある程度の説明力を持っているとしても、回答者の数が少なすぎるなどの理由で、母集団の推測にとっては有意でないことがある。

ここで行う検定は、説明力が少なくとも 0 ではない（決定係数 $R^2 \neq 0$）といってよいかどうかの検定であり、下の計算式で算出される F 値を検定統計量として利用する。F 値は、ランダムな誤差に対して独立変数による説明が何倍の予測力を持っているか、という分散比を表すことになる。

$$F = \frac{R^2}{(1-R^2)/(n-2)}$$

したがって、F 値が十分に大きく、ランダム誤差の何倍もの説明力が認められるならば、回帰線は母集団についても説明力を持つとみなされる（$R^2 \neq 0$）。計算式から分かるように、F 値が大きくなるのは、決定係数 $R^2$ が大きいときと、標本の回答者数 n が大きいときである。

遅刻回数の例では、決定係数 $R^2$ が 0.438 で、回答者数 n が 15 であったので、F 値は次のような値をとり、アルバイト時間はランダム誤差に比べて 10 倍程度の説明力をもつ。

$$F = \frac{0.438}{(1-0.438)/(15-2)} = 10.13$$

確率表にあてはめると、このような F 値がまったくの偶然に出現する確率（有意確率）は、わずかに 0.7%程度しかない（p＝0.007）。したがって、標準的に 5%を有意水準とするならば、この回帰線は十分に統計的に有意であり、母集団についてもこの回帰線で物事を考えることに統計的な意味があると認められる。

**（練習）**

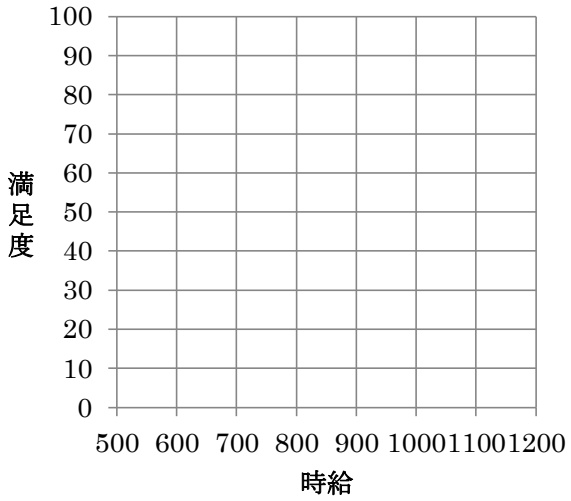
回帰分析の結果が以下のとおりであった場合、具体的にどういう意味が読み取れるか、穴埋めしてみよう。

○飲食店のアルバイト店員 50 名に対するアンケートデータを用いた回帰分析

・従属変数は「アルバイトへの満足度（100 点満点)」
・独立変数は「アルバイトの時給」
・回帰分析の結果、定数項 $\alpha$ ＝-55.8、回帰係数 $\beta$ ＝0.13
・調整済み決定係数 $R^2$＝0.113
・F 値を検定統計量とした検定の結果、有意確率 p＝0.0098

↓

回帰分析で求められた最適な回帰式は、$\hat{Y}$＝＿＿＿＿＿＿＿＿＿＿＿＿で、回帰線をおよそのグラフで図示すると、下のようになる。具体的には、たとえば時給が 700 円のときの満足度は＿＿＿＿点と予測されるのに対して、時給が 900 円ならば、満足度＿＿＿＿点と予測される。




また、この結果から、時給によってアルバイトの満足度は、およそ＿＿＿＿%説明することができることがわかる。この 50 名のアンケート結果から、時給でアルバイトの満足度がある程度説明できると一般化してよいかというと、偶然このような結果が得られた確率（有意確率）が＿＿＿＿%なので、統計的に有意な結果と｛いえる・いえない｝。

「回帰分析（2）：SPSS で実践」

## ■SPSS でやってみよう

前回は、回帰分析の考え方について学習した。今回は SPSS を操作して、実際に回帰分析の結果を出力しながら、一通りの手続きを経験しよう。

---

回帰分析の操作

---

①メニューから、分析➜回帰➜線型
②説明したい変数（Y）を［従属変数］、説明に使う変数（X）を［独立変数］欄へ移動
（②’　質的変数を独立変数にする場合は、あらかじめダミー変数に変換すること）


③ＯＫボタン

---

**モデル要約** ❷

| モデル | R | R2 乗 | 調整済み R2 乗 | 推定値の標準誤差 |
|---|---|---|---|---|
| 1 | .662[a] | .438 | .394 | 6.97320 |

a. 予測値: (定数)、work アルバイト時間。

**分散分析[a]** ❸

| モデル | | 平方和 | 自由度 | 平均平方 | F 値 | 有意確率 |
|---|---|---|---|---|---|---|
| 1 | 回帰 | 491.868 | 1 | 491.868 | 10.115 | .007[b] |
| | 残差 | 632.132 | 13 | 48.626 | | |
| | 合計 | 1124.000 | 14 | | | |

a. 従属変数 late 遅刻回数
b. 予測値: (定数)、work アルバイト時間。

**係数[a]**

| ❶ モデル | | 標準化されていない係数 B | 標準誤差 | 標準化係数 ベータ | t 値 | 有意確率 ❹ |
|---|---|---|---|---|---|---|
| 1 | (定数) | 3.009 | 2.841 | | 1.059 | .309 |
| | work アルバイト時間 | .095 | .030 | .662 | 3.180 | .007 |

a. 従属変数 late 遅刻回数

読み取るポイント
❶最適な回帰式の α 、 β
❷調整済み決定係数
❸全体的な説明力の検定結果

（重回帰分析の場合）
❹各独立変数の
影響力の検定結果

---

■独立変数が複数の場合の回帰分析

ここまでは、独立変数が 1 つの場合の回帰分析を扱ったが、一般的には複数の独立変数を用いた回帰分析がよく行われる。独立変数が複数の場合を**重回帰分析**（multiple regression analysis）と呼ぶこともあるが、回帰分析といえばふつうは重回帰分析のことである。

独立変数が複数ある場合の回帰式は、次のようにどんどん独立変数の効果を足し合わせていく形で表現される。

$$\hat{Y} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots$$

これはつまり、独立変数の値が 1 増加することは、(他の要素とは関係なく）常に一定の影響力で従属変数の値に作用する､つまり傾きが一定である､という考え方を踏襲している。図形として視覚化することはできないが、多次元空間の散布図の中に 1 本の最適な線を通して、常に一定の法則が働いていることを主張しようとしていることを意味する。重回帰分析の回帰係数（$\beta_1$、$\beta_2$、$\beta_3$、……）は、とくに偏回帰係数と呼ぶこともある。

具体的には、たとえば遅刻回数 Y を、アルバイト時間 $X_1$、通学時間(分)$X_2$、睡眠時間 $X_3$ で説明しようとする重回帰分析では、次のような形で最適な回帰式が析出される。

$$\hat{Y} = 10.21 + 0.22 X_1 + 0.04 X_2 - 1.31 X_3$$

この場合、アルバイトが 1 時間増えるごとに遅刻が 0.22 回増え、同様に通学時間が 1 分長いごとに 0.04 回遅刻が増える。睡眠時間が 1 時間長いごとに遅刻は 1.31 回減る。すべての独立変数が 0 ならば、遅刻は 10.21 回と予測される。独立変数が 1 つの場合と、読み方はまったく同じである。

分析の手続きもほぼ同様であり、以下の 4 点にまとめられる。

(1) もっともよい回帰式を定める。
　　（最小二乗法で、α、$\beta_1$、$\beta_2$、……$\beta_k$の値を定める）
(2) その回帰式は、どのくらいよい式であるかを評価する。
　　（[調整済み] 決定係数によって、説明力を算出する）
(3) 母集団についても、同様の回帰式を定める価値があるかどうか判断する。
　　（全体的な説明力を F 値によって検定する）
(4) 母集団についても、各独立変数を説明に用いる価値があるか、個別に判断する。
　　（それぞれの独立変数の影響を t 値によって検定する）

4 つ目の手続だけが重回帰分析に独自のものである。回帰式全体の説明力について検定するだけではなく、1 つ 1 つの独立変数が従属変数を説明するために有効に働いているかどうか、それぞれの影響について検定する。つまり、それぞれの回帰係数$\beta_1$、$\beta_2$、……について、母集団でも一定の影響力がある（$\beta \neq 0$）といってよいかどうかを検定する。

この検定は、t 値と呼ばれる検定統計量を用い、統計分析ソフトでは対応する有意確率が同時に示される。ここでの有意確率は、つまり、回帰分析で示されている回帰係数がまったくの偶然の産物である確率なので、この確率が一定の値よりも低ければ、偶然ではなく母集団でもその独立変数に一定の影響力があるとみなしてよいことになる。

**（練習）**

1. 実際の全国調査（JGSS-2000）から抽出した 30 代男性のデータを用いて、月給を従属変数、年齢を独立変数とする(月給の違いを年齢で説明する)回帰分析を実行してみよう。

→読み取るポイント
①最適な回帰式

②調整済み決定係数

③全体的な説明力の検定結果

2. 独立変数を、「年齢」「勤続年数」「中 3 の頃の成績」の 3 つとして、月給を説明する重回帰分析を実行してみよう。

→読み取るポイント
①最適な回帰式

②調整済み決定係数

③全体的な説明力の検定結果

④各独立変数の影響力の検定結果

実習用データ　　http://www2.itc.kansai-u.ac.jp/~tyasuda/

「回帰分析（3）：発展」

**■質的変数を独立変数にする場合：ダミー変数**

回帰分析の独立変数は量的変数であることが基本である。しかし、質的変数も工夫をすれば独立変数として分析に用いることができる。社会調査データには質的変数が多いので、この応用は重要である。

回帰分析で質的変数を用いる場合には、ダミー変数に変換した上で用いる。ダミー変数とは、0 か 1 のどちらかの値しか取らない変数のことである。たとえば、性別という変数を独立変数に用いたいときには、図 1 のように男性を 1 とするダミー変数（男性ダミー）か、女性を 1 とするダミー変数（女性ダミー）のいずれかにリコーディングし、そのダミー変数を回帰分析に用いる。

| | 元の変数 | | 男性ダミー | | 女性ダミー |
|---|---|---|---|---|---|
| 男性 | 1 | → | 1 | または | 0 |
| 女性 | 2 | → | 0 | | 1 |

図 1　性別のダミー変数

ダミー変数を用いた回帰式の読み取りは簡単である。たとえば、Y が遅刻回数、$X_1$ が学年、$X_2$ が男性ダミーの重回帰分析で次のような回帰式が求められたとする。

$$\hat{Y} = 2.0 + 3.9X_1 + 2.2X_2$$

この場合、男子学生は女子学生に比べて 2.2 回多く遅刻することが読み取れる。

性別は 2 つのグループしかない質的変数であったが、3 つ以上のグループ（カテゴリー）がある質的変数の場合はどうすればよいのだろうか。たとえば、学生が所属する学部を独立変数に用いたいが、学部は文学部、法学部、工学部、医学部と 4 種類あるとする。この場合、図 2 のように 3 つのダミー変数を作成し、これらすべてを独立変数に用いた回帰分析を行えばよい。

| | 元の変数 | | 文学部ダミー | 法学部ダミー | 工学部ダミー |
|---|---|---|---|---|---|
| 文学部 | 1 | → | 1 | 0 | 0 |
| 法学部 | 2 | → | 0 | 1 | 0 |
| 工学部 | 3 | → | 0 | 0 | 1 |
| 医学部 | 4 | → | 0 | 0 | 0 |

図 2　学部のダミー変数

もう 1 つ医学部ダミーが必要ではないかと思うかもしれないが、4 つ目のダミー変数は

不要である。なぜならば、文学部ダミー、法学部ダミー、工学部ダミーの値がいずれも 0 である回答者は、自動的に医学部なので、3 つのダミー変数さえあれば 4 つの学部のどれに所属しているか区別できるからである。一般に k 個のグループ（カテゴリー）の質的変数の内容は、1 つ少ない k－1 個のダミー変数で表すことができる。ここでは、医学部ダミーを除いているが、医学部ダミーを分析に加えて他の 3 つのダミー変数のうち 1 つを分析から除いてもかまわない。

このようなダミー変数の回帰係数は、省略したカテゴリー（ここでは医学部）と比べて、当該のカテゴリーであることがもたらす影響力を表すことになる。たとえば、文学部ダミーの回帰係数が 1.2 であれば、それは「医学部と比べて」文学部の方が 1.2 回だけ遅刻が多いと予測されることを意味する。「文学部以外と比べて」という意味にはならないので注意しよう。

したがって、ダミー変数を省略したカテゴリーは、比較の基準になるという意味で意外と重要な意味を持つ。このようなカテゴリーを**参照カテゴリー**（reference category）と呼ぶ。いまの例の場合には、医学部が参照カテゴリーである。

参照カテゴリーは、分析者が結果の読み取りやすさを考えて選ぶもので、決まった選び方はない。しかし、次の 2 点に注意する必要がある。1 つは、参照カテゴリーは内容のはっきりとしたグループでなければならない。たとえば、「その他」というグループを参照カテゴリーにすると、何と比べているのか分からなくなるので避ける。もう 1 つの注意点として、参照カテゴリーのグループに属する回答者は、ある程度人数が多いことが望ましい。あまりに人数が少ないグループを基準にして比較をすると、分析結果が不安定なものになってしまう。

SPSS では、「他の変数への値の再割り当て」という機能を使って、ダミー変数を作成することができる。ややめんどうであるが、質的変数を回帰分析に活用するためには必要な作業である。

**■標準化回帰係数**

重回帰分析では、いったいどの独立変数が一番影響力をもつのか、といったことに関心が向くことがある。単純に回帰係数を比べるだけでは、この疑問に答えることはできない（独立変数の単位が違うため）。たとえば、1 日の歩行量が 1 歩増えるごとに、体重が 1.5g 減り（$\beta_1$＝－1.5）、1 ヶ月にジムに通う回数が 1 回増えるごとに、体重が 500g 減る（$\beta_2$＝－500）としても、ジムに通う回数の方が体重に強く影響するということにはならない。

このような比較をおこなうときに有効なのが、**標準化回帰係数**（standardized regression coefficient）である。標準化回帰係数は、通常の回帰係数に独立変数と従属変数の標準偏差の比を掛け合わせたもので、すべての変数を標準得点にしたとき（標準偏差を 1 に調整したとき）、独立変数が 1 点増えることが従属変数を何点増やすことになるのかを表す。つまり、すべての変数の単位（ばらつきの程度）をそろえることで、各独立変数の効果を比較できるようにしている。

たとえば、体重の標準偏差が 10,000g（10kg）、歩行量の標準偏差が 2,000 歩、ジムに通う回数の標準偏差が 3 回だったとすると、それぞれの独立変数の標準化回帰係数は、次のようになるので、歩行量の方が影響の規模が大きいことが分かる。

$$\beta_1^* = -1.5 \times \frac{2000}{10000} = -0.3\,, \qquad \beta_2^* = -500 \times \frac{3}{10000} = -0.15$$

SPSS では「標準化係数ベータ」という列に、自動的に各独立変数の標準化回帰係数が表示されるので、とくに苦労なくこの値を用いることができる。

**■独立変数の出し入れ**

重回帰分析では､同じ独立変数でも､他にどのような独立変数を投入したのかによって、回帰係数が変わってくる。たとえば、性別（男性ダミー）と年齢で月給の額を説明しようとしたとき､男性ダミーの回帰係数が 10 万だったとする(男性の方が月給が 10 万円高い)。しかし、これに加えて、正規雇用ダミーを独立変数に加えると、男性ダミーの回帰係数が 5 万に減少したりすることがある。

これは、重回帰分析が「ワンセットの独立変数で」従属変数を説明する回帰線を求めるからである。つまり､「性別と年齢だけで説明しなさい」と言われれば、性別の効果が大きいという説明をせざるをえないが､「正規雇用という原因で説明してもいいよ」と言われれば、性別が男性だからという理由で説明するよりも、正規雇用のおかげで月給が高いと説明する方が適切だ、という解答を回帰分析は示してくれる。

このようなことが起こるのは、そもそも性別と正規雇用の間に強い関連性があるからである（男性の方が正規雇用が多い)。独立変数群の中に関連性の強い変数の組み合わせがあるときには、その回帰係数に注意して、一方の変数を出し入れしてみると、回帰分析の結果がどう変わるかを観察してみよう｡扱っている現象に対する理解が深まるはずである(見せかけの関係や媒介関係といった統計的な現象を熟知していれば、理解はより深まる)。

また、このことからもわかるように、回帰分析はあくまで分析者が提示したモデル（変数間の因果関係の枠組み）の中で最適な答えを出しているにすぎないことを、忘れないようにしなければならない。回帰分析が「正しい因果関係」を示してくれるわけではない。分析者が想定した因果関係の枠組みの中で、各独立変数の具体的な影響力の大きさ（回帰係数）について最適解を知らせてくれるだけである。したがって、回帰分析はある程度そのメカニズムが理解できている社会現象について、より詳細な情報を得るために用いるべきである。

**■分散分析と一般線型モデル**

テキストでは回帰分析といっしょに、分散分析、一般線型モデルといった分析技法が紹介されている。これらは、非常に関連の深い技法なので、簡単にその意味を解説しておこう。

**分散分析**（analysis of variance; ANOVA）は、ふつう、質的変数を独立変数として、そのグループの間で従属変数の平均値に差があるといってよいかどうかを検定するための技法として用いられる。たとえば、文学部と法学部と社会学部の間で、大学満足度の平均値に差があるかどうかを検定したりする。

これがなぜ回帰分析と関係するのかといえば、独立変数のグループによって従属変数の平均値が違うかどうかを調べることと、独立変数が従属変数の値にどのように影響する

か（回帰係数の規模はどうか）を調べることは、結局同じことだからである。独立変数が質的変数だったり量的変数だったりの違いがあるように見えるが、回帰分析で質的変数をダミー変数に変換して扱えることからもかわるように、この違いは数学的には問題にならない。そのため、回帰分析と分散分析を区別せずに、1 つの同じものとして、**一般線型モデル**（general linear model; 一般線形モデルとも書く）と呼ぶことがある。それぞれの独立変数の影響の有無に関心を集中させるなら分散分析になり、独立変数の影響の程度に関心を向けるならば回帰分析になる。

実際に、SPSS による回帰分析の出力の中には「分散分析表」と名前が付いている部分があり、F 値による全体的な説明力の検定がおこなわれている。分散分析は、この F 値の算出にもっとこだわりを見せる。つまり、全体的な説明力だけでなく、各独立変数を加えることが説明力に与える影響や、独立変数の組み合わせを考えることが説明力を上げるかどうか（たとえば、性別と年齢それぞれの影響だけでなく、20 代男性といった組み合わせに意味があるかどうかなど）を調べたりする。このようなこだわりを見せる際には、回帰分析のようにそれぞれの独立変数の影響力についてその規模（回帰係数）まで見ようとするよりも、それぞれの影響力の有無に絞って検定結果（F 値）に注視する方がよい。それが分散分析である。

**＜参考文献＞**

岩井紀子・保田時男，2007，『調査データ分析の基礎』 有斐閣.

村瀬洋一・高田洋・廣瀬毅士，2007，『SPSS による多変量解析』 オーム社.

小田利勝，2007，『ウルトラ・ビギナーのための SPSS による統計解析入門』 プレアデス出版.