

★SPSSで重回帰分析を行う

重回帰分析は、①従属変数を予測するための式を作る ②従属変数との関係が強い独立変数群を探す ときに使われます。

例えば、ある店の売り上げがあったとします。その額は、いろいろな要因によって変動するはずです。

要因をいくつか挙げてみると、

- 1.店の立地条件
- 2.品揃え
- 3.店員の接客態度
- 4.周囲の人口
- 5.店の規模 etc.

これらの要因から売上額を推定する式をたてるのが重回帰分析です。

重回帰分析で導かれる式は、次のようになります。

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_nx_n$$

上の例をふまえて書き直すとこうなります。

売上額 = 定数 + 立地条件の重み × 立地条件 + 品揃えの重み × 品揃え + … + 店の規模の重み × 店の規模

重みとはいったい何でしょう？簡単に言うと、どれくらい重視するかの目安です。立地条件が売上額に大きく影響しているなら、その重みは大きくなります。

仮に、重みをつけて式を書き直してみます。

売上額 = 定数 + (-1000) × 立地条件 + 3000 × 品揃え + 1500 × 接客態度 + 1800 × 人口 + 2500 × 規模

ある店が、立地条件…駅から徒歩10分 品揃え…50種類 接客態度…良(ここでは便宜的に3点とする) 人口…250世帯 規模…400㎡ という要因を持っていたとすると、その店の売上額は次のようになります。

$$\text{売上額} = \text{定数} + (-1000) \times 10 + 3000 \times 50 + 1500 \times 3 + 1800 \times 250 + 2500 \times 400$$

$$\rightarrow \text{売上額} = \text{定数} + 1594500$$

定数は、想定した要因以外の要因からなる数値です。定数が1000000なら、この店の売上額は2594500円となります。

もし定数・品揃え・接客態度・人口・規模が全く同じで、立地条件が駅から徒歩5分の店があれば、売上額は2599500円となります。

この重みのことを**偏回帰係数**と呼びます。非標準化係数とも言います。

もし偏回帰係数の数が一つしかなければどうなるでしょう？もちろん式は次のようになります。

$$y = a + bx$$

これは単純な回帰直線の式になります。

では、これからSPSSを使った手順の説明に移ります。

以下のようなデータがあったとします。

| | 給料(円) | 勤続年数(年) | 仕事量(個/時) | 欠勤日数(日) | 特殊免許 |
|---|--------|---------|----------|---------|------|
| 1 | 213000 | 8 | 261 | 2 | 1 |
| 2 | 192000 | 5 | 188 | 4 | 0 |
| 3 | 199000 | 7 | 167 | 2 | 0 |
| 4 | 241000 | 13 | 244 | 0 | 0 |
| 5 | 175000 | 2 | 122 | 6 | 0 |
| 6 | 220000 | 10 | 238 | 2 | 0 |
| 7 | 208000 | 4 | 253 | 1 | 1 |
| 8 | 184000 | 3 | 196 | 2 | 0 |

| | | | | | |
|----|--------|----|-----|---|---|
| 9 | 203000 | 7 | 201 | 5 | 0 |
| 10 | 245000 | 12 | 278 | 0 | 1 |
| 11 | 188000 | 6 | 166 | 8 | 0 |
| 12 | 216000 | 10 | 256 | 7 | 0 |
| 13 | 200000 | 6 | 193 | 3 | 1 |
| 14 | 173000 | 1 | 132 | 1 | 0 |
| 15 | 187000 | 3 | 170 | 2 | 0 |
| 16 | 191000 | 3 | 181 | 1 | 1 |
| 17 | 219000 | 9 | 240 | 1 | 1 |
| 18 | 189000 | 2 | 194 | 8 | 1 |
| 19 | 236000 | 11 | 255 | 0 | 1 |
| 20 | 179000 | 1 | 143 | 0 | 1 |

これはある工場の従業員の給料と、従業員の属性です。
これを重回帰分析にかけ、給料を予測する式を得ようと思います。
架空のデータなので、細かいことはあんまり気にしてはいけません。

これをSPSSの分析にかけます。

“分析”→“回帰”→“線型”を選択します。



この場合yにあたるものは給料です。

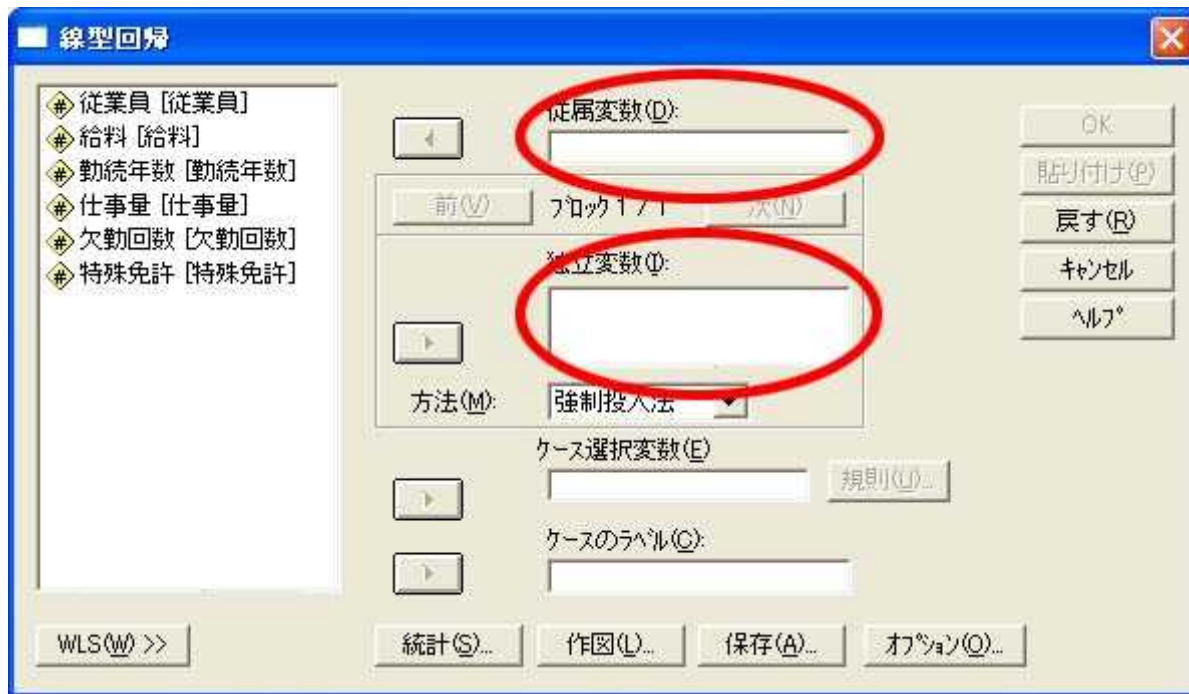
上の従属変数には「給料」を入れます。

「勤続年数」「仕事量」「欠勤」「免許」は、
「給料」を説明する変数なので、独立変数に入れます。

次に方法を選択するのですが、目的により使い分けます。

強制投入法…独立変数を全て使って式をたてる

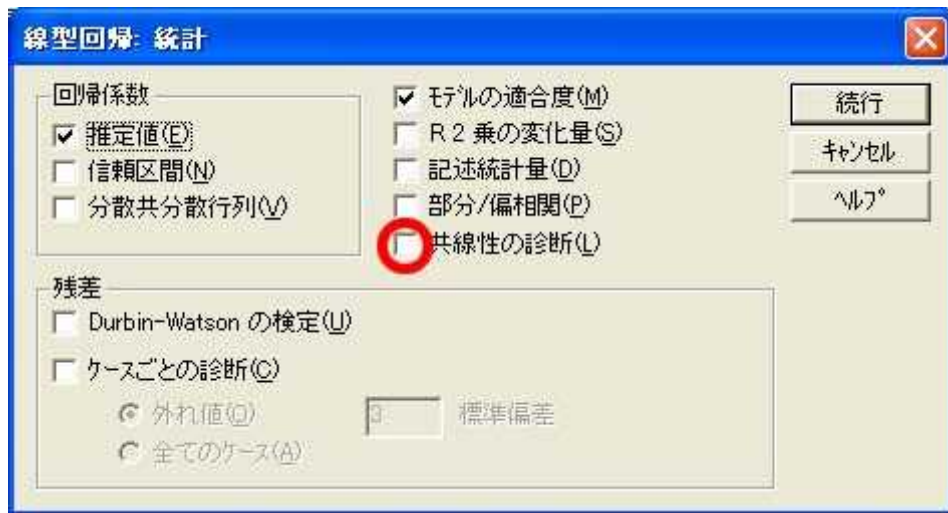
ステップワイズ法…独立変数を徐々に増やして式をたてる



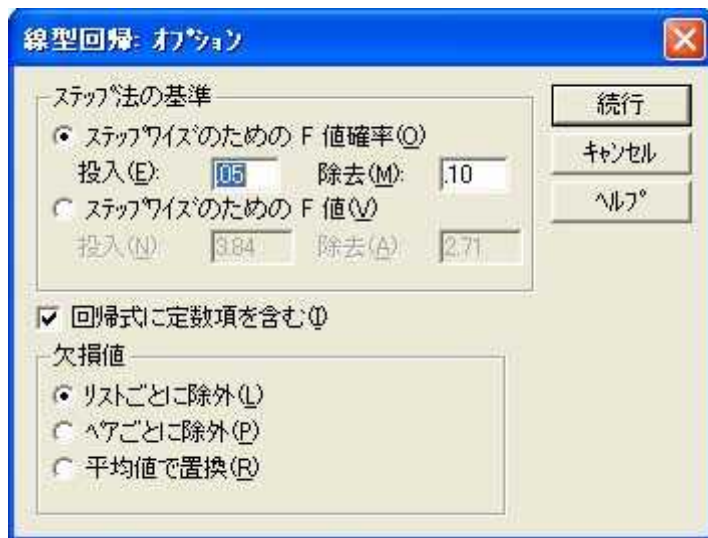
独立変数が確定している場合は強制投入法を、独立変数が未確定ならステップワイズ法を使ってください。もちろん他の方法を用いても構いません。ここではステップワイズ法を用いたとして話を進めます。

次に“統計”をクリックします。

ここでは“共線性の診断”にチェックを入れます。共線性については後で説明します。“続行”をクリックしてください。



次は“オプション”です。



ステップワイズ法では、独立変数を徐々に加えて最適な式を導き出すものです。その変数の加え方をここでは設定しますが、初期の設定でとりあえずは大丈夫です。分析をしたら独立変数が一つになってしまった、などのときにこれらの条件を緩和してください。

どうしても駄目なら、ステップワイズ法を諦め、強制投入法などを使ってください。“続行”をクリックしてください。

これで準備が整ったので、“OK”をクリックします。

結果の見方を解説します。

投入済み変数または除去された変数^a

| モデル | 投入済み変数 | 除去された変数 | 方法 |
|-----|--------|---------|---|
| 1 | 勤続年数 | | ステップワイズ法 (基準: 投入する F の確率 $\leq .050$ 、除去する F の確率 $\geq .100$)。 |
| 2 | 特殊免許 | | ステップワイズ法 (基準: 投入する F の確率 $\leq .050$ 、除去する F の確率 $\geq .100$)。 |
| 3 | 仕事量 | | ステップワイズ法 (基準: 投入する F の確率 $\leq .050$ 、除去する F の確率 $\geq .100$)。 |
| 4 | 欠勤回数 | | ステップワイズ法 (基準: 投入する F の確率 $\leq .050$ 、除去する F の確率 $\geq .100$)。 |

a. 従属変数 給料

まずこんな表が出力されます。

ここでは、

モデル1…給料を説明するのに勤続年数だけを使ったモデル

モデル2…給料を説明するのに勤続年数と特殊免許を使ったモデル

モデル3…給料を説明するのに勤続年数と特殊免許と仕事量を使ったモデル

モデル4…給料を説明するのに勤続年数と特殊免許と仕事量と欠勤回数を使ったモデル

ということです。

通常は一番下にあるモデルを使います。この場合、モデル4を採用することになります。

このデータの場合、全ての独立変数が給料にかかわるという結果になっています。

元のデータによっては、全ての変数が投入されないこともあります。

強制投入法では、有無を言わず全ての変数を投入します。

モデル集計

| モデル | R | R2乗 | 調整済み R2乗 | 推定値の 標準誤差 |
|-----|-------------------|------|-------------|--------------|
| 1 | .942 ^a | .887 | .881 | 7384.32010 |
| 2 | .973 ^b | .946 | .940 | 5245.25732 |
| 3 | .980 ^c | .961 | .954 | 4618.96855 |
| 4 | .986 ^d | .972 | .964 | 4038.87205 |

- a. 予測値: (定数)、勤続年数。
- b. 予測値: (定数)、勤続年数、特殊免許。
- c. 予測値: (定数)、勤続年数、特殊免許、仕事量。
- d. 予測値: (定数)、勤続年数、特殊免許、仕事量、欠勤回数。

次にこの表です。前述したとおり、モデル4以外は無視しても構いません。

この表で重要なのはR2乗です。これを**重決定係数**と言います。

これはモデルのあてはまりの良さを示します。

重決定係数が良い値かそうでないかは、次の分散分析表を見なければなりません。

分散分析^e

| モデル | | 平方和 | 自由度 | 平均平方 | F値 | 有意確率 |
|-----|----|-----------|-----|-----------|---------|-------------------|
| 1 | 回帰 | 7.742E+09 | 1 | 7.742E+09 | 141.987 | .000 ^a |
| | 残差 | 981507300 | 18 | 54528183 | | |
| | 全体 | 8.724E+09 | 19 | | | |
| 2 | 回帰 | 8.256E+09 | 2 | 4.128E+09 | 150.041 | .000 ^b |
| | 残差 | 467716313 | 17 | 27512724 | | |
| | 全体 | 8.724E+09 | 19 | | | |
| 3 | 回帰 | 8.382E+09 | 3 | 2.794E+09 | 130.966 | .000 ^c |
| | 残差 | 341357928 | 16 | 21334870 | | |
| | 全体 | 8.724E+09 | 19 | | | |
| 4 | 回帰 | 8.479E+09 | 4 | 2.120E+09 | 129.948 | .000 ^d |
| | 残差 | 244687311 | 15 | 16312487 | | |
| | 全体 | 8.724E+09 | 19 | | | |

- a. 予測値: (定数)、勤続年数。
- b. 予測値: (定数)、勤続年数、特殊免許。
- c. 予測値: (定数)、勤続年数、特殊免許、仕事量。
- d. 予測値: (定数)、勤続年数、特殊免許、仕事量、欠勤回数。
- e. 従属変数: 給料

これも例によってモデル4だけを見れば大丈夫です。

この表によると、有意確率は.000ですから、

この分析によって得られた式は非常にあてはまりが良いと言えます。

係数^a

| モデル | | 非標準化係数 | | 標準化係数 | t | 有意確率 | 共線性の統計量 | |
|-----|------|------------|----------|-------|--------|------|---------|-------|
| | | B | 標準誤差 | ベータ | | | 許容度 | VIF |
| 1 | (定数) | 170000.739 | 3217.044 | | 52.844 | .000 | | |
| | 勤続年数 | 5349.473 | 448.938 | .942 | 11.916 | .000 | 1.000 | 1.000 |
| 2 | (定数) | 165565.955 | 2505.001 | | 66.094 | .000 | | |
| | 勤続年数 | 5324.993 | 318.942 | .938 | 16.696 | .000 | 1.000 | 1.000 |
| | 特殊免許 | 10189.647 | 2357.940 | .243 | 4.321 | .000 | 1.000 | 1.000 |
| 3 | (定数) | 151130.531 | 6328.506 | | 23.881 | .000 | | |
| | 勤続年数 | 4244.136 | 525.485 | .747 | 8.077 | .000 | .286 | 3.502 |
| | 特殊免許 | 6673.495 | 2529.607 | .159 | 2.638 | .018 | .674 | 1.485 |
| | 仕事量 | 111.157 | 45.675 | .241 | 2.434 | .027 | .248 | 4.025 |
| 4 | (定数) | 154458.444 | 5700.066 | | 27.098 | .000 | | |
| | 勤続年数 | 4055.461 | 465.980 | .714 | 8.703 | .000 | .278 | 3.601 |
| | 特殊免許 | 4883.696 | 2330.903 | .116 | 2.095 | .054 | .607 | 1.649 |
| | 仕事量 | 116.884 | 40.008 | .254 | 2.922 | .011 | .248 | 4.039 |
| | 欠勤回数 | -919.920 | 377.888 | -.115 | -2.434 | .028 | .842 | 1.188 |

結果のメインがこの表です。
これもモデル4に注目します。

青い枠内が、式となります。すなわち、
給料 = 154458.444(定数)

$$\begin{aligned}
 &+ 4055.461 \times \text{勤続年数} \\
 &+ 4883.696 \times \text{特殊免許} \\
 &+ 116.884 \times \text{仕事量} \\
 &+ (-919.920) \times \text{欠勤回数}
 \end{aligned}$$

になります。ということは、勤続5年、特殊免許なし、
仕事量200個/時、欠勤回数1回の従業員は、
154458.444 + 2077.305 + 0 + 23376.8 - 919.920
= 178992.629円が給料の期待値となります。

a. 従属変数 給料

偏回帰係数はわかりましたが、係数の大きさは単純に比較することはできません。

つまり特殊免許の係数と仕事量の係数を比較して、特殊免許の方が比重が大きいとは言えないのです。なぜなら単位が違うからです。

特殊免許は有るか無いか、つまり1か0かです。一方仕事量は大体150~250という数値です。

では実際、特殊免許と仕事量では、どちらが給料に占める割合が大きくなっているのでしょうか？そこで緑の枠をみてください。

これは標準化係数(β、または標準偏回帰係数)といって、標準得点に直したものです。つまり同じものさしで測っているわけです。

表から、特殊免許の標準化係数βは0.116、仕事量の標準化係数は0.254ですから、仕事量のウエイトは特殊免許よりも大きくなっています。

次は除去された変数という表がありますが、これは無視して構いません。

最後に、共線性の表を見ます。

共線性の診断^a

| モデル | 次元 | 固有値 | 条件指標 | 分散の比率 | | | | |
|-----|----|-----------|--------|-------|------|------|------|------|
| | | | | (定数) | 勤続年数 | 特殊免許 | 仕事量 | 欠勤回数 |
| 1 | 1 | 1.858 | 1.000 | .07 | .07 | | | |
| | 2 | .142 | 3.620 | .93 | .93 | | | |
| 2 | 1 | 2.414 | 1.000 | .03 | .04 | .06 | | |
| | 2 | .453 | 2.307 | .03 | .14 | .84 | | |
| | 3 | .132 | 4.272 | .94 | .82 | .09 | | |
| 3 | 1 | 3.384 | 1.000 | .00 | .01 | .02 | .00 | |
| | 2 | .471 | 2.680 | .00 | .03 | .62 | .00 | |
| | 3 | .138 | 4.959 | .08 | .28 | .07 | .00 | |
| | 4 | 7.741E-03 | 20.907 | .91 | .68 | .30 | 1.00 | |
| 4 | 1 | 3.851 | 1.000 | .00 | .00 | .01 | .00 | .01 |
| | 2 | .690 | 2.362 | .00 | .00 | .23 | .00 | .30 |
| | 3 | .360 | 3.272 | .00 | .08 | .30 | .00 | .27 |
| | 4 | 9.127E-02 | 6.496 | .13 | .25 | .19 | .01 | .42 |
| | 5 | 7.740E-03 | 22.306 | .87 | .66 | .26 | .99 | .00 |

a. 従属変数 給料

共線性とは、独立変数間に相関があることを言います。要因同士に相関があると、いろいろ問題が起きます。予測式が狂う可能性を秘めていることとなります。

共線性の診断では、条件指標にまず注目します。これの大きいところに共線性の可能性があります。条件指標が大きいところを見つけたら、横を見ます。この表の場合、勤続年数と仕事量の値が高いです。つまり、勤続年数と仕事量は相関していると言えます。

共線性が見つかったら、その変数を除去するなどしてから、もう一度分析をしたり、相関行列に切り替えたりといった対処をしてください。

これで重回帰分析の解説を終わります。

[戻る](#)