2) 相関係数

2 つの変数の相関の程度を表す指標として相関係数がある。相関係数は次式の通り定義されるけれども、やや難しいのでここでは理解する必要はない (興味ある読者は章末の「統計学からの補足」を参照)。

$$\text{相関係数} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

相関係数は Excel では「=CORREL(範囲 1, 範囲 2)」関数で簡単に計算できる。相関係数は -1 から 1 の間の値を取る。1 に近いほど正の相関が高く、-1 に近いほど負の相関が高く、0 に近い場合は無相関である。

練習問題 4.2

C51 に 1 人あたり県民所得と第 2 次産業比率の相関係数を計算してみよう。

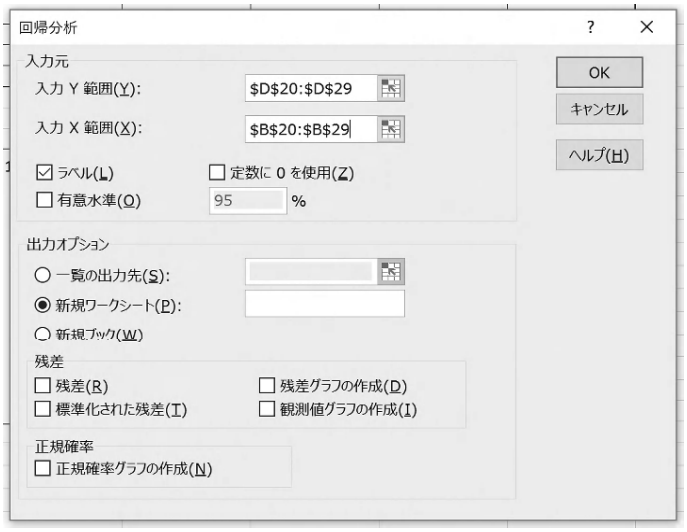
散布図を見ると、これらの変数間にはかなり高い正の相関があると予想できるが、計算結果は約 0.3~~5~~4 で、無相関となってしまう。なぜだろうか。

練習問題 4.1 で作成した散布図を見ると、東京都のデータだけが離れた位置にプロットされていた。このようなデータを**外れ値**という。外れ値があると相関係数が低下するので、東京のデータを除いて相関係数を計算してみると 0.76 となり、かなり高い正の相関があることが分かる (東京のデータを一時的に削除すれば確認できる。確認したら元に戻すボタンで元に戻そう)。

3) 散布図を利用したデータの分類

散布図は 2 変数の相関関係を分析するだけではなく、データのグルーピングにも役立てることができる。しばしば使われる方法は、2 つの変数の平均と比較して上か下かによって 4 つのグループに分ける方法である。しかし、データが明確に 4 つのグループに分かれるとは限らず、分類に統計学的な根拠はないので、その解釈は分析者のセンスに委ねられる。

図 5-2　回帰分析の入力画面



3) 分析結果の読み方

数字がたくさん並んでいるが、さしあたり以下の点に注目すればよい。

重決定 R2

重決定 R2 は決定係数といわれる数値で、回帰式のあてはまり具合を表す指標である。分析結果は 0.826031 なので、この式で y の 82 % を説明できていることを意味する。

有意 F

有意 F は回帰直線の説明力が有意かどうかを示す F 値の有意確率を示している。分析結果は 0.0007 であり、有意水準 0.05（5 %）より小さいので、回帰直線は統計的に有意であるといえる。

係数

係数は回帰係数といわれる数値である「切片」は回帰式の α、「年齢」は β を表し、回帰式は以下の通りとなる。 これにより、年齢が分かれば大卒女性の平均年収が推計できる (単位は千円)。

概要

回帰統計	
重相関 R	0.908863
重決定 R2	0.826031
補正 R2	0.801179
標準誤差	500.0743
観測数	9

分散分析表

	自由度	変動	分散		有意 F
回帰	1	8311747	8311747	33.23711	0.000688
残差	7	1750520	250074.3		
合計	8	10062267			

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	1968.34	567.3389	3.469425	0.010414	626.7965	3309.883
年齢	74.439	12.91186	5.765164	0.000688	43.9073	104.9707

回帰係数は、その符号と値の大きさに注目する。正の場合は説明変数と被説明変数が比例し、負の場合は反比例する。この式の場合、説明変数「年齢」の係数の符号がプラスなので、給与は年齢とともに増加することを意味し、妥当な結果と言える。係数の大きさは、年齢が 1 歳大きくなるごとに年間給与が約 7 万 4 千円ずつ増加することを意味している。

$$\text{年間給与} = 1968.\cancel{298}\,34 + 74.44 \times \text{年齢}$$

t 値

「t」は t 値と呼ばれる統計値である。t 値は、回帰係数の統計的な有効性を表す。t 値の絶対値が「臨界値」以上であれば、説明変数として統計的に有意であると判断できる。臨界値は t 分布表で調べることができるが、Excel では「=TINV(有意確率, 自由度)」で計算できる。自由度はデータ数マイナス変数の数で、単回帰分析の場合は (データ数 − 2) である。この分析の場合、有意確率 5 %、自由度 7 なので、臨界値は 2.36 となる。データ数が大きくなると臨界値は小さくなり、データ数が無限大の場合 1.96 となるから、データ数が大きいときには t 値が 2 以上かどうかが大

6. 重回帰分析

1) 重回帰分析とは

単回帰分析は被説明変数を 1 つの説明変数で説明しようとするものであるが、経済現象は複雑なものなので、2 変数の関係だけで分析できるとは限らず、2 つ以上の説明変数で回帰分析をすることの方が一般的である。これを重回帰分析という。

x_1 と x_2 の 2 つの説明変数を用いた重回帰分析は、以下のような重回帰式を想定していることになる。回帰係数 β_1, β_2 は偏回帰係数と呼ばれることもある。というのは、β_1 は、x_2 が一定の時 (変化しない時)、x_1 だけが変化した時の y の変化量を表すからである。

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Excel の「データ分析」機能を使って重回帰分析を行うには、「入力X範囲」で複数行にわたってデータ範囲を選択するだけでよいので、単回帰分析の方法を理解していればすぐに実行できるだろう。ただし、ここではもう少し新しい知識を学んでから分析に進むことにする。

2) 2 次関数による回帰分析

賃金データを使った回帰分析で、女性の年収は単回帰分析によって 80 % 程度の説明力がある回帰式を求めることができたが、男性の年収は単回帰分析ではやや説明力が弱かった (~~補正~~ R2 が 0.~~48~~、有意 F は 0.02)。それは、男性の年収が、直線というよりは 50 歳代前半をピークにその後減少していく山形をしているからである。このようなデータの場合、次のような 2 次関数 (2 乗項を含む関数) に当てはめると説明力の高い回帰式を得られることがある。

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

重回帰分析

・「データ分析」を実行し、「入力Y範囲」は年間給与額を指定し、「入力X範囲」は勤続年数と学歴ダミーの範囲 (A20:C47) を指定して (その他は前と同じ)、OK をクリックする。結果は以下の通りとなる

概要

回帰統計	
重相関 R	0.941927
重決定 R2	0.887227
補正 R2	0.872518
標準誤差	672.4178
観測数	27

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	3	81815892	27271964	60.31676	4.71E-11
残差	23	10399352	452145.7		
合計	26	92215244			

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	1814.326	275.4777	6.586109	1.02E-06	1244.457	2384.195
勤続年数	112.8145	11.8147	9.548656	1.81E-09	88.37391	137.255
高校卒	617.8222	316.9808	1.949084	0.063584	-37.9025	1273.547
大学卒	2854.267	316.9808	9.004541	5.32E-09	2198.542	3509.991

「高校卒」以外は

結果の考察

重決定係数は 0.89、自由度調整済み決定係数は 0.87、有意 F は 0.000 で、推計された重回帰式は十分な説明力があるといえる。偏回帰係数の P 値はいずれも 0.05(5%) を下回っているので、すべての説明変数は統計的に有意であるといえる。

回帰式は以下の通りである。高校卒ダミー変数の係数は 617.82 なので、中学卒に比べて年収が約 61 万 8 千円高いことになる。また、大学卒ダミー変数の係数は 2854.27 なので、中学卒より約 285 万 4 千円高いことになる。

$$\begin{aligned} \text{年間給与額} &= 1814.33 + 112.81 \times \text{勤続年数} \\ &\quad +617.82 \times \text{高校卒} + 2854.27 \times \text{大学卒} \end{aligned}$$

おり、「高校卒」は0.1（10%）を下回っているので

中学卒はダミー変数が両方とも 0 であり、年間給与総額は次式で表される。

$$\text{中学卒の年間給与額} = 1808.627945 + 116.9618729 \times \text{勤続年数}$$

高校卒は高校卒ダミー変数が 1、大学卒ダミー変数は 0 で、式は以下の通り。

$$\begin{aligned}\text{高校卒の年間給与額} &= 1808.627945 + 116.9618729 \times \text{勤続年数} \\ &\quad + 828.1111111 \\ &= 2636.739056 + 116.9618729 \times \text{勤続年数}\end{aligned}$$

大学卒は高校卒ダミー変数が 0、大学卒ダミー変数 1 で、式は以下の通り。

$$\begin{aligned}\text{大学卒の年間給与額} &= 1808.627945 + 116.9618729 \times \text{勤続年数} \\ &\quad + 2905.555556 \\ &= 4714.183501 + 116.9618729 \times \text{勤続年数}\end{aligned}$$

4) 重回帰分析をする際の注意点

・**多重共線性**

説明変数に用いる複数のデータの相関が高いとき、回帰分析の推定の精度が悪くなる。このような場合、多重共線性が生じているという。たとえば、賃金は年齢と勤続年数のどちらとも関係あるが、年齢と勤続年数は相関が高く、多重共線性が生じる可能性がある。

・**系列相関 (自己相関)**

時系列データを用いて回帰分析をする際に、前期のデータが今期のデータと相関が高い場合、このデータに系列相関 (自己相関) があるという。系列相関があると分析結果が実際よりも見かけ上良くなる。

多重共線性や系列相関を検定する方法については第 6 節で簡単に触れるが、詳しくは計量経済学で学んで欲しい。

練習問題 6.4

日本では急速な勢いで少子化が進んでおり、大きな社会問題となっている。少子化を食い止めるためにはどのような対策を実施すればよいだろうか。適切