

重回帰分析

1. 重回帰分析の寄与率 R がわかる
2. 重回帰分析と単回帰分析の比較がわかる
3. 重回帰分析の推定区間の式が導出できる
4. 偏相関係数が導出できる
5. 重回帰分析の多重共線性がわかる
6. 重回帰分析は単位に影響されない理由がわかる(その 1)
7. 重回帰分析は単位に影響されない理由がわかる(その 2)
8. 偏回帰係数に関する検定と推定がよくわかる
9. 変数増減法がよくわかる

問 1：重回帰分析の寄与率 R は、 $0 \leq R \leq 1$ になる理由を説明せよ。

問 2：重回帰分析の寄与率 R が 0 になる時は、どんな場合か？具体的なデータ例を示せ。

問 3：重回帰分析の寄与率 R が 1 になる時は、どんな場合か？具体的なデータ例を示せ。

多変量解析はすべて数式で導出できます。導出過程から本質を理解しましょう。

【1】寄与率 R が $0 \leq R \leq 1$ な理由がわかる

(1) 寄与率 R とは

どの教科書でも書いていますが、回帰平方和 S_R と総平方和 S_T の比を寄与率 R と定義します。

$$R = \frac{S_R}{S_T} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

y_i ：実測値、 $\hat{y}_i$ ：回帰直線上の点、 $\bar{y}$ ：平均

なお、データの構造式は、

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

となり、2乗和を取ると、それぞれの項が

$$S_T (\text{総平方和}) = S_R (\text{回帰平方和}) + S_{er} (\text{回帰残差平方和})$$

に分解できます。

この証明は、関連記事で解説しているので、ご確認ください。

【関連記事】平方和の分解と分散分析ができる(重回帰分析)

<https://qcplanets.com/method/multi-regression/anova-basic/>

(2) 単回帰分析の寄与率 R が $0 \leq R \leq 1$ になる理由がわかる

まず、単回帰分析について解説します。あとで、重回帰分析の場合の証明でも有効ですが、別の方法でも証明ができます。

単回帰分析の寄与率 R が $0 \leq R \leq 1$ になり理由は

QC プラネッツ単回帰分析プレミアム勉強プリント

https://qcplanets.com/wp-content/uploads/2025/09/simple_linear_regression_qcplanets_texts.pdf

No.1 【必読】相関係数や寄与率が 1 以上にできない理由がわかる

で確認ください。

この関連記事を書いたあとで、気づいたのですが、これから解説する、重回帰分析の寄与率の場合の証明の方が、簡単に証明できます。

(3) 重回帰分析の寄与率 R が $0 \leq R \leq 1$ になる理由がわかる

寄与率の定義に戻ると、

$$R = \frac{S_R}{S_T} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$S_T (\text{総平方和}) = S_R (\text{回帰平方和}) + S_{er} (\text{回帰残差平方和})$$

を使って、証明できます。

① 寄与率 $R \geq 0$ を証明する

まず、寄与率 R の式の分母分子に注目すると、

$$\bullet (\text{分子}) = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = (\hat{y}_1 - \bar{y})^2 + (\hat{y}_2 - \bar{y})^2 + \dots + (\hat{y}_n - \bar{y})^2$$

とすべての 0 に 2 乗がついているので実数である限り、すべて 0 以上になります。よって、(分子) ≥ 0

●同様に(分母)も 0 に 2 乗がついているので実数である限り、すべて 0 以上になります。よって、(分母) ≥ 0

つまり、寄与率 R は 0 より大きい分母分子の比なので、

$R \geq 0$ は明らかですね。

②寄与率 $R \leq 1$ を証明する

次に、寄与率 R は

$$R = \frac{S_R}{S_T} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

でかつ、

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

ですから、この式の(両辺)を $\sum_{i=1}^n (y_i - \bar{y})^2$ で割ります。

$$1 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} + \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

となり、(右辺)の第1項をよく見ると $R = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$ なので、

$1 = R + \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$ となります。また、(右辺)の第2項は分母、分子をよく見ると0以上ですね。

よって、

$$1 - R = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \geq 0$$

より、 $R \leq 1$ が証明できます。以上まとめると、

$$0 \leq R \leq 1$$

が証明できました。

【2】寄与率 R が0になる条件

(1) データの構造式、平方和から導出

データの構造式から平方和を求める式を見れば、わかりやすいです。平方和を求める式を再掲すると、

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \text{ ですね。}$$

寄与率 R が0は、 $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = 0$ です。
 Σ の中はすべて2乗ですから、 Σ 全体=0にするには、
すべての()の中が0であることが必須ですね。

つまり、

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = (\hat{y}_1 - \bar{y})^2 + (\hat{y}_2 - \bar{y})^2 + \dots + (\hat{y}_n - \bar{y})^2$$

のすべての()の中が0なので、

$$\bullet \hat{y}_1 - \bar{y} = 0$$

$$\bullet \hat{y}_2 - \bar{y} = 0$$

...

$$\bullet \hat{y}_n - \bar{y} = 0$$

となります。整理すると、

$$\hat{y}_1 = \hat{y}_2 = \dots = \hat{y}_n = \bar{y}$$

(すべての回帰データと平均値が一致する場合)

(2) 寄与率 R が0になる回帰式

説明変数が n 個ある、重回帰分析において、回帰式は、

$$\hat{y}_i = \bar{y} + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

と表記できます。

$\hat{y}_1 = \hat{y}_2 = \dots = \hat{y}_n = \bar{y}$ を満たすには、

$$\beta_1 = \beta_2 = \dots = \beta_n = 0 \text{ が必要ですね。}$$

傾きが 0 なデータの場合、寄与率 $R = 0$ となるとわかります。

(3) 寄与率 R が 0 になる実例

具体的なデータを上げてみましょう。意外と難しいですが、1 例は下表のデータです。

-	x1	x2	y	S_{11}	S_{1y}	S_{12}	S_{22}	S_{2y}	S_{yy}
-	1	2	3	6.25	-2.5	7.5	9	-3	1
-	2	3	2	2.25	0	3	4	0	0
-	3	4	1	0.25	0.5	0.5	1	1	1
-	4	6	1	0.25	-0.5	0.5	1	-1	1
-	5	7	2	2.25	0	3	4	0	0
-	6	8	3	6.25	2.5	7.5	9	3	1
合計	21	30	12	17.5	0	22	28	0	4

回帰平方和 S_R は

$$S_R = \beta_1 S_{1y} + \beta_2 S_{2y} = 0$$

となるので、寄与率 R は

$$R = \frac{S_R}{S_T} = \frac{0}{4} = 0$$

となり、確かに、寄与率 $R=0$ になりました。

【3】寄与率 R が 1 になる条件がわかる

(1) データの構造式、平方和から導出

データの構造式から平方和を求める式を見れば、わかりやすいです。平方和を求める式を再掲すると、

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

ですね。

寄与率 R が 1 は回帰残差平方和に注目して

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = 0$$

Σ の中はすべて 2 乗ですから、 Σ 全体=0 にするには、すべての $(y_i - \hat{y}_i)$ の中が 0 であることが必須ですね。

つまり、

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2 = 0$$

のすべての $(y_i - \hat{y}_i)$ の中が 0 なので、

$$\bullet y_1 - \hat{y}_1 = 0$$

$$\bullet y_2 - \hat{y}_2 = 0$$

...

$$\bullet y_n - \hat{y}_n = 0$$

となります。整理すると、

$$y_1 = \hat{y}_1, y_2 = \hat{y}_2, \cdots, y_n = \hat{y}_n,$$

と実測値と回帰直線上の理論値がすべて一致する場合は寄与率 $R=1$ になります。

(2) 寄与率 R が 1 になる実例

実測値と回帰直線上の理論値がすべて一致するデータを探しましょう。

-	x1	x2	y(=3x ₁ +2x ₂ -1)	S ₁₁	S _{1y}	S ₁₂	S ₂₂	S _{2y}	S _{yy}
-	1	2	6	6.25	33.75	7.5	9	40.5	182.25
-	2	3	11	2.25	12.75	3	4	17	72.25
-	3	4	16	0.25	1.75	0.5	1	3.5	12.25
-	4	6	23	0.25	1.75	0.5	1	3.5	12.25
-	5	7	28	2.25	12.75	3	4	17	72.25
-	6	8	33	6.25	33.75	7.5	9	40.5	182.25
合計	21	30	12	17.5	96.5	22	28	122	533.5

上表の y をみると

$$y=3 \times x_1 + 2 \times x_2 - 1$$

で値を代入しており、誤差 ε は含んでいません。そうすると、確かに

回帰平方和 S_R は

$$S_R = \beta_1 S_{1y} + \beta_2 S_{2y} = 3 \times 96.5 + 2 \times 122 = 533.5 = S_{yy} = S_T$$

となり、 $S_R = S_T$ となるので、寄与率 R は

$$R = \frac{S_R}{S_T} = 1 \text{ になりました。}$$

以上、「重回帰分析の寄与率 R がわかる」を解説しました。

多重共線性の話につながるテーマです。

【1】重回帰分析と単回帰分析を比較

(1) 目的変数データは同じで、説明変数の種類を変えて寄与率を比較

寄与率を比較するために、目的変数 z は同じとし、
説明変数は

● 単回帰分析は x のみ

● 重回帰分析は x, y とし、

重回帰分析と単回帰分析の x データも同じとし、 y の値の差によって、寄与率がどう変化するかを考えます。

イメージを下表に書きます。

単回帰分析		重回帰分析		
x	z	x	y	z
1	3	1	??	3
4	4	4	??	4
2	4	2	??	4
5	7	5	??	7
4	7	4	??	7
2	5	2	??	5
寄与率	A	寄与率	B	A、Bどちらがおおきい？

(2) 公式を比較

まず、重回帰分析と単回帰分析において、それぞれの

● 「データの構造式」 ● 「回帰直線」 ● 「回帰直線の傾き」 ● 「平方和」 ● 「寄与率」

について公式を確認しましょう。公式は暗記ではなく導出できます。関連記事で導出過程を確認しましょう。

★単回帰分析の復習

【関連記事】回帰分析と相関係数をマスターする

<https://qcplanets.com/method/regression/basic/>

★重回帰分析の復習

【関連記事】重回帰分析の回帰式が導出できる

<https://qcplanets.com/method/multi-regression/equation/>

★公式を比較

	単回帰分析	重回帰分析
データの構造式	$(z_i - \bar{z}) = (\hat{z}_i - \bar{x}) + (z_i - \hat{z}_i)$	
回帰直線	$z = \alpha + \beta x$	$z = \eta + \gamma x + \delta y$
傾き	$\beta = \frac{S_{xz}}{S_{xx}}$	(*)
回帰平方和	$S_{R1} = \frac{S_{xz}^2}{S_{xx}}$	$S_{R2} = \gamma S_{xz} + \delta S_{yz}$
総平方和	$S_T = S_{zz}$	
寄与率R	$R1 = \frac{S_{R1}}{S_T}$	$R2 = \frac{S_{R2}}{S_T}$

上表の(*)は、下の連立方程式を満たす解が傾き γ, δ となります。

$$S_{xx}\gamma + S_{xy}\delta = S_{xz}$$

$$S_{xy}\gamma + S_{yy}\delta = S_{yz}$$

(3) 寄与率を比較

上の表を使って、重回帰分析と単回帰分析の寄与率の差を比較しましょう。ところで、寄与率 R を求める際、 $R = \frac{S_R}{S_T}$ ですが、単回帰分析も重回帰分析も目的変数のデータが同じ場合は、どちらも、総平方和 $S_T = S_{ZZ}$ なので、寄与率を求める分子の S_R について比較します。

★寄与率を比較(≡回帰平方和を比較)

回帰平方和 S_R で比較すると、

●単回帰分析： $S_{R1} = \frac{S_{xz}^2}{S_{xx}}$

●重回帰分析： $S_{R2} = \gamma S_{xz} + \delta S_{yz}$

比較します。

$$S_{R2} - S_{R1} = (\gamma S_{xz} + \delta S_{yz}) - \frac{S_{xz}^2}{S_{xx}} \quad (\text{式 1})$$

ここで、

● $\gamma S_{xz} = \beta S_{xx}$

● $S_{R1} = \frac{S_{xz}^2}{S_{xx}} = \beta^2 S_{xx}$

の関係式を代入すると(式 1)は

$$(\text{式 1}) = (\gamma S_{xz} + \delta S_{yz}) - \beta^2 S_{xx} = \beta(\gamma - \beta) S_{xx} + \delta S_{yz}$$

となります。

$$S_{R2} - S_{R1} = \beta(\gamma - \beta) S_{xx} + \delta S_{yz}$$

の正負で寄与率の大小が決まります。でも、大小比較しにくい式ですね。

基本、 S_{xx} , S_{yz} は正でそれ以外の傾きの値の正負で大小関係が決まりますが、式だけではわかりにくいので、実例を挙げて大小関係を見ましょう。

【2】重回帰分析と単回帰分析で寄与率が等しい場合

(1) データ事例

下表を再掲しますが、**重回帰分析と単回帰分析で寄与率を等しくするには、 y のデータ値をどうすればよいかわかりますか？**

単回帰分析		重回帰分析		
x	z	x	y	z
1	3	1	??	3
4	4	4	??	4
2	4	2	??	4
5	7	5	??	7
4	7	4	??	7
2	5	2	??	5
寄与率	A	寄与率	B	A、Bどちらがおおきい？

答えは、

x, y のデータを同じにすれば寄与率は等しくなる！

つまり、

単回帰分析		重回帰分析		
x	z	x	y	z
1	3	1	1	3
4	4	4	4	4
2	4	2	2	4
5	7	5	5	7
4	7	4	4	7
2	5	2	2	5
寄与率	A	寄与率	B	A、Bどちらがおおきい？

とする場合、重回帰分析と単回帰分析の寄与率は等しくなります。

(2) 寄与率が等しくなる理由

x と y が完全に一致するので、

$$\bullet S_{xy}=S_{xx}$$

$$\bullet S_{yz}=S_{xz}$$

$$\bullet S_{yy}=S_{xx}$$

となります。また、重回帰分析から傾き γ, δ を求める連立方程式は

$$S_{xx}\gamma + S_{xy}\delta = S_{xz}$$

$$S_{xy}\gamma + S_{yy}\delta = S_{yz}$$

が

$$S_{xx}\gamma + S_{xx}\delta = (\gamma + \delta)S_{xx} = S_{xz}$$

$$S_{xy}\gamma + S_{yy}\delta = (\gamma + \delta)S_{xx} = S_{xz}$$

という 1 つの式になってしまいます。一方、重回帰分析の回帰平方和 S_R は

$$S_{R2} = \gamma S_{xz} + \delta S_{yz} \text{ より、}$$

$$S_{R2} = \gamma S_{xz} + \delta S_{yz} = \gamma S_{xz} + \delta S_{xz} = (\gamma + \delta)S_{xz} = (\gamma + \delta)^2 S_{xx}$$

となります。単回帰分析の回帰平方和 S_R は

$$S_{R1} = \frac{S_{xz}^2 - \frac{\beta^2 S_{xx}^2}{S_{xx}}}{S_{xx}} = \beta^2 S_{xx}$$

となります。

$$S_{R2} = (\gamma + \delta)^2 S_{xx}$$

$$S_{R1} = \beta^2 S_{xx}$$

で回帰直線はともに、

$$\bullet \text{重回帰直線} : z = \eta + \gamma x + \delta y = \eta + (\gamma + \delta)x$$

$$\bullet \text{単回帰直線} : z = \alpha + \beta x$$

を比較すると

$$\eta = \alpha, \gamma + \delta = \beta \text{ より、}$$

$$S_{R2} = S_{R1}$$

となり、寄与率 R も等しくなります。

(3) 実際に寄与率を計算してみる

分散分析と寄与率を計算すると、下表のように確かに、重回帰分析と単回帰分析の寄与率は一致します。

-	S(単回帰)	S(重回帰)
R	8.33	8.33
e	5.67	5.67
T	14	14
寄与率	0.595	0.595

【3】重回帰分析の方が単回帰分析より寄与率が等しい場合

(1) 一般には重回帰分析の方が寄与率は高くなる

$$S_{R2} - S_{R1} = \beta(\gamma - \beta)S_{xx} + \delta S_{yz}$$

の正負で寄与率の大小が決まります。でも、大小比較しにくい式ですね。

基本は、説明変数 y は x と関係ないデータが入るため、それなりの平方和があります。

目的変数 z を同じとすると、総平方和 S_T は同じのまま、回帰平方和 S_R 説明変数があるだけ平方和が加算されるので、**一般には重回帰分析の方が寄与率は高くなります。**

1 つ実データを入れてみましょう。

(2) データ事例

説明変数 y に下表のような値を入れてみましょう。ただし、説明変数 $x \neq y$ とします。

x	y	z
1	3	3
4	2	4
2	4	4
5	4	7
4	5	7
2	6	5
18	24	30

分散分析、寄与率を計算し、単回帰分析と重回帰分析で比較しましょう。

単回帰分析 (x,z)	平方和	自由度	重回帰分析 (x,y,z)	平方和	自由度
回帰R	8.33	1	回帰R	13.04	2
残差e	5.67	4	残差e	0.96	3
計T	14	5	計T	14	5

確かに、回帰平方和 S_R は、
単回帰分析 8.33 から
重回帰分析 13.04 に
増えていますね。

$$S_{R2} - S_{R1} = \beta(\gamma - \beta)S_{xx} + \delta S_{yz}$$

の正負で寄与率の大小が決まります。

上の式では、減少する要素より増加する要素が大きいから、回帰平方和 S_R は増加したと考えられます。

説明変数 x_2 にいろいろ値を入れてみると、ほぼ重回帰分析の回帰平方和 S_R の方が値は大きいです。

【4】単回帰分析の方が重回帰分析より寄与率が大きい場合

(1) 事例はないかも？

$$S_{R2} - S_{R1} = \beta(\gamma - \beta)S_{xx} + \delta S_{yz}$$

の正負で寄与率の大小が決まります。この値が負になるデータ例が見つかりませんでした。

以上、「重回帰分析と単回帰分析の比較がわかる」を解説しました。

【1】推定区間の式(その1)

(1) 推定区間の式を紹介

目的変数 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$

の信頼度 $(100 - \alpha)\%$ の推定区間は、

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p \pm t(n - p - 1, \alpha) \sqrt{\left(1 + \frac{1}{n} + \frac{D^2}{n-1}\right) V_e}$$

で与えられる。Dはマハラビノス距離

(2) 推定区間の式をよくみると

推定区間の式は基本 (平均) $\pm t(\text{自由度}, \alpha) \sqrt{\frac{V_e}{n_e}}$ なので、

● (平均) $= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$

● $t(n - p - 1, \alpha) = t(\text{自由度}, \alpha) \rightarrow t$ 分布に従うことは理解できる

● $\sqrt{\left(1 + \frac{1}{n} + \frac{D^2}{n-1}\right) V_e}$ って何なん？

【2】導出に必要な関係式を導出(その1)

重回帰分析の回帰式の作り方については、関連記事で確認ください。

【関連記事】重回帰分析の回帰式が導出できる

<https://qcplanets.com/method/multi-regression/equation/>

(1) 平方和 S_{ij} と S^{ij}

重回帰分析の回帰式は

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$ で表現しますが、

実測値と回帰直線上の予測値の間の誤差が最小になる条件から回帰式の傾きや y 切片を求めます。

具体的には最小 2 乗法を使うと、次の式ができます。行列を使いますが、

$$\begin{pmatrix} S_{11} & S_{12} & \dots & S_{1p} \\ S_{21} & S_{22} & \dots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \dots & S_{pp} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} S_{1y} \\ S_{2y} \\ \vdots \\ S_{py} \end{pmatrix}$$

で表現できますね。なお、S は各成分の平方和です。逆行列を使って、 β_i の各値を計算します。つまり、

$$\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} S^{11} & S^{12} & \dots & S^{1p} \\ S^{21} & S^{22} & \dots & S^{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S^{p1} & S^{p2} & \dots & S^{pp} \end{pmatrix} \begin{pmatrix} S_{1y} \\ S_{2y} \\ \vdots \\ S_{py} \end{pmatrix}$$

となります。ここで、 S^{ij} は逆行列の i 行 j 列目の値で、添え字を上側とします。

(2) 平方和 S_{ij} と S^{ij} の関係式

互いに逆行列の関係なので、実際に計算してみましょう。

$$\begin{pmatrix} S_{11} & S_{12} & \dots & S_{1p} \\ S_{21} & S_{22} & \dots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \dots & S_{pp} \end{pmatrix} \begin{pmatrix} S^{11} & S^{12} & \dots & S^{1p} \\ S^{21} & S^{22} & \dots & S^{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S^{p1} & S^{p2} & \dots & S^{pp} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} = E(\text{単位行列})$$

実際に

$$\bullet (i,j)=(1,1): S_{11}S^{11} + S_{12}S^{21} + \dots + S_{1p}S^{p1}=1$$

$$\bullet (i,j)=(1,2): S_{11}S^{12} + S_{12}S^{22} + \dots + S_{1p}S^{p2}=0$$

...

$$\bullet (i,j)=(1,p): S_{11}S^{1p} + S_{12}S^{2p} + \dots + S_{1p}S^{pp}=0$$

と j について計算し、これを i についてどんどん具体的に計算しましょう。

$$\bullet (i,j)=(2,1): S_{21}S^{11} + S_{22}S^{21} + \dots + S_{2p}S^{p1}=0$$

$$\bullet (i,j)=(2,2): S_{21}S^{12} + S_{22}S^{22} + \dots + S_{2p}S^{p2}=1$$

...

$$\bullet (i,j)=(2,p): S_{21}S^{1p} + S_{22}S^{2p} + \dots + S_{2p}S^{pp}=0$$

...

$$\bullet (i,j)=(p,p): S_{p1}S^{1p} + S_{p2}S^{2p} + \dots + S_{pp}S^{pp}=1$$

となりますね。

★展開式から規則性を探そう！

いっぱい式を展開しましたが、規則性を探すと

● $i=j$ の時は 1、それ以外は 0 デルタ関数っぽい条件式ができる

● (左辺)の展開式は $\sum_{k=1}^p S_{ik}S^{kj}$ でまとめられる

シンプルな式でまとめると

$$\sum_{k=1}^p S_{ik}S^{kj} = 1 \quad (i = j)$$

$$\sum_{k=1}^p S_{ik}S^{kj} = 0 \quad (i \neq j)$$

となります。この関係式をあとで使います。

丁寧に導出しましたが、ここを手抜きすると、後の導出の理解ができなくなります。下ごしらえは丁寧にやりましょう。

【3】傾き β_j の期待値が導出できる(その 1)

(1) 傾き β_j の式

傾きを計算する行列の式を再掲します。

$$\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} S^{11} & S^{12} & \dots & S^{1p} \\ S^{21} & S^{22} & \dots & S^{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S^{p1} & S^{p2} & \dots & S^{pp} \end{pmatrix} \begin{pmatrix} S_{1y} \\ S_{2y} \\ \vdots \\ S_{py} \end{pmatrix}$$

ここで、j 列にある β_j を求める式を抜き出すと

$$\beta_j = S^{j1} S_{1y} + S^{j2} S_{2y} + \dots + S^{jk} S_{ky} + \dots + S^{jp} S_{py}$$

ですね。

(2) 傾き β_j の期待値

上の式の期待値 $E[\beta_j]$ は、

$$E[\beta_j] = E[S^{j1} S_{1y} + S^{j2} S_{2y} + \dots + S^{jk} S_{ky} + \dots + S^{jp} S_{py}] = (\text{式 1})$$

ですね。ここで、

分布関数に従うのは誤差を含む y であり、 x_i だけで作られた値はすべて定数扱いとするので、 S^{jk} は定数とし、期待値 E の外に出せます。

$$(\text{式 1}) = S^{j1} E[S_{1y}] + S^{j2} E[S_{2y}] + \dots + S^{jk} E[S_{ky}] + \dots + S^{jp} E[S_{py}] = \sum_{k=1}^p S^{jk} E[S_{ky}] = (\text{式 2})$$

とまとめることができます。

(3) 関係式を代入して傾き β_j の期待値を計算

ここで、 S_{ky} は、傾き β_j を求める行列の式から

$$\begin{pmatrix} S_{11} & S_{12} & \dots & S_{1p} \\ S_{21} & S_{22} & \dots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \dots & S_{pp} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} S_{1y} \\ S_{2y} \\ \vdots \\ S_{py} \end{pmatrix}$$

$$S_{ky} = \beta_1 S_{k1} + \beta_2 S_{k2} + \dots + \beta_p S_{kp}$$

と計算できるので、(式 2) に代入します。

$$(\text{式 2}) = \sum_{k=1}^p S^{jk} E[S_{ky}] = \sum_{k=1}^p S^{jk} E[\beta_1 S_{k1} + \beta_2 S_{k2} + \dots + \beta_p S_{kp}] = (\text{式 3})$$

期待値 E の [] の中は、変数 x_i について値なので、定数扱いとして $E[]$ の外に出せます。よって(式 3)は

$$(\text{式 3}) = \sum_{k=1}^p S^{jk} S_{k1} \beta_1 + \sum_{k=1}^p S^{jk} S_{k2} \beta_2 + \dots + \sum_{k=1}^p S^{jk} S_{kp} \beta_p = (\text{式 4})$$

となります。

上の「②導出に必要な関係式を導出」で

$$\sum_{k=1}^p S_{ik} S^{kj} = 1 \quad (i = j)$$

$$\sum_{k=1}^p S_{ik} S^{kj} = 0 \quad (i \neq j)$$

を(式 4)に代入すると、

(式 4)の中の $\sum_{k=1}^p S^{jk} S_{kj} \beta_j$ のみ ($i=j$) が

$$\sum_{k=1}^p S^{jk} S_{kj} \beta_j = 1 \times \beta_j$$

となり、それ以外 ($i \neq j$) は

$$\forall \sum_{k=1}^p S^{jk} S_{kj} \beta_j = 0 \text{ になります。}$$

よって、まとめると

$$(\text{式 4}) = \sum_{k=1}^p S^{jk} S_{kj} \beta_j = \beta_j$$

となります。

$$E[\beta_j] = \beta_j$$

とわかりましたが、当たり前ですね。

ですが、分散の計算に必要な前座でもあるので、期待値を丁寧に導出しました。

【4】傾き β_j の分散が導出できる

(1) 大事な関係式を再掲

$$\beta_j = S^{j1} S_{1y} + S^{j2} S_{2y} + \dots + S^{jk} S_{ky} + \dots + S^{jp} S_{py}$$

$$\sum_{k=1}^p S_{ik} S^{kj} = 1 \quad (i = j)$$

$$\sum_{k=1}^p S_{ik} S^{kj} = 0 \quad (i \neq j)$$

となります。この関係式をあとで使います。

(2)分散の式を作る

では行きましょう。傾き β_j の分散は

$$V[\beta_j] = V[S^{j1} S_{1y} + S^{j2} S_{2y} + \cdots + S^{jk} S_{ky} + \cdots + S^{jp} S_{py}] = (\text{式 5})$$

(式 5)を展開していきます。

$$(\text{式 5}) = V[\sum_{k=1}^p S^{jk} \sum_{i=1}^n (x_{ik} - \bar{x}_k)(y_i - \bar{y})] = (\text{式 2})$$

(式 2)の中の

$$\sum_{k=1}^n (x_{ik} - \bar{x}_k)(y_i - \bar{y}) = \sum_{k=1}^n (x_{ik} - \bar{x}_k)y_i - \bar{y} \sum_{k=1}^n (x_{ik} - \bar{x}_k) = \sum_{k=1}^n (x_{ik} - \bar{x}_k)y_i$$

$$(\text{第 2 項の} \sum_{k=1}^n (x_{ik} - \bar{x}_k) = 0)$$

これを(式 2)に代入します。

$$(\text{式 2}) = V[\sum_{k=1}^p S^{jk} \sum_{i=1}^n (x_{ik} - \bar{x}_k)y_i] = V[\sum_{i=1}^n (\sum_{k=1}^p S^{jk} (x_{ik} - \bar{x}_k) y_i)] = (\text{式 3})$$

と x についての項だけでまとめます。

(式 3)について、 x についての項は定数扱いなので、2乗にして $V[\]$ の外に出せます。

$$(\text{式 3}) = \sum_{i=1}^n (\sum_{k=1}^p S^{jk} (x_{ik} - \bar{x}_k) y_i)^2 V[y_i] = \sigma^2 \sum_{i=1}^n (\sum_{k=1}^p S^{jk} (x_{ik} - \bar{x}_k) y_i)^2 = (\text{式 4})$$

$$(V[y_i] = \sigma^2)$$

(式 4)は

$$(\text{式 4}) = \sigma^2 \sum_{i=1}^n \sum_{k=1}^p S^{jk} (x_{ik} - \bar{x}_k) \sum_{l=1}^p S^{jl} (x_{il} - \bar{x}_l) = \sigma^2 \sum_{k=1}^p \sum_{l=1}^p S^{jk} S^{jl} \sum_{i=1}^n (x_{ik} - \bar{x}_k)(x_{il} - \bar{x}_l) = (\text{式 5})$$

(式 5)で、 $\sum_{i=1}^n (x_{ik} - \bar{x}_k)(x_{il} - \bar{x}_l) = S_{kl}$ なので、代入すると

$$(\text{式 5}) = \sigma^2 \sum_{k=1}^p \sum_{l=1}^p S^{jk} S^{jl} S_{kl} = \sigma^2 \sum_{k=1}^p S^{jk} \delta_{jk} = \sigma^2 S^{jj}$$

出ました！まとめると、

$$V[\beta_j] = \sigma^2 S^{jj}$$

【5】 傾き β_j の共分散が導出できる(研究中) </h2>

実は、これはまだよくわかっていませんが、

$$V[\beta_j] = \sigma^2 S^{jj} = \text{Cov}(\beta_j, \beta_j) = \sigma^2 S^{jj}$$

と見立てると

$$\text{Cov}(\beta_j, \beta_j) = \sigma^2 S^{jj}$$

とできる！

ちょっと強引だし、「何で？」となるので、わかり次第報告しますが、この式を使わせていただきます。

【6】 推定区間の式が導出できる

だいぶ準備かけましたが、いよいよ推定区間の式を導出しましょう。

(1) 推定区間の式

目的変数 y は

$$y = \bar{y} + \beta_1(x_1 - \bar{x}_1) + \beta_2(x_2 - \bar{x}_2) + \cdots + \beta_p(x_p - \bar{x}_p) + \varepsilon$$

ですね。この値の分散 $V[y]$ を計算しましょう。マハラビノス距離が出て来ます。

(2) 分散 $V[y]$ を計算

$$V[y] = V[\bar{y} + \beta_1(x_1 - \bar{x}_1) + \beta_2(x_2 - \bar{x}_2) + \cdots + \beta_p(x_p - \bar{x}_p) + \varepsilon]$$

分散を展開すると

$$\begin{aligned} V[y] &= V[\bar{y}] + V[\beta_1(x_1 - \bar{x}_1) + \beta_2(x_2 - \bar{x}_2) + \cdots + \beta_p(x_p - \bar{x}_p)] + V[\varepsilon] \\ &+ 2\text{Cov}(\bar{y}, \beta_1(x_1 - \bar{x}_1) + \beta_2(x_2 - \bar{x}_2) + \cdots + \beta_p(x_p - \bar{x}_p)) \\ &+ 2\text{Cov}(\beta_1(x_1 - \bar{x}_1) + \beta_2(x_2 - \bar{x}_2) + \cdots + \beta_p(x_p - \bar{x}_p), \varepsilon) \\ &+ 2\text{Cov}(\bar{y}, \varepsilon) \end{aligned}$$

ここで、 $\bar{y}, \beta_1(x_1 - \bar{x}_1) + \beta_2(x_2 - \bar{x}_2) + \cdots + \beta_p(x_p - \bar{x}_p), \varepsilon$ は互いに独立と仮定するので、共分散 Cov はすべて 0 になります。また、

$$\bullet V[\bar{y}] = V\left[\frac{y}{n}\right] = \frac{\sigma^2}{n}$$

$$\bullet V[\varepsilon] = \sigma^2$$

を代入すると

$$\begin{aligned} V[y] &= \frac{\sigma^2}{n} + V[\beta_1(x_1 - \bar{x}_1) + \beta_2(x_2 - \bar{x}_2) + \cdots + \beta_p(x_p - \bar{x}_p)] + \sigma^2 \\ &= (\text{式 6}) \text{となります。} \end{aligned}$$

あとは、

$$V[\beta_1(x_1 - \bar{x}_1) + \beta_2(x_2 - \bar{x}_2) + \cdots + \beta_p(x_p - \bar{x}_p)] = (\text{式 7})$$

を計算すればよいことになります。やってみましょう。

(式 7) を計算して、 $V[y] = (\text{式 6})$ をまとめましょう。

(3) 分散 $V[y]$ を計算

(式 7) において、まず $(x_i - \bar{x}_i)$ は定数扱いなので、分散 V 、共分散 Cov の外側に出せます。分散の加法性を使って、めっちゃ展開します。

$$\begin{aligned} (\text{式 7}) &= V[\beta_1(x_1 - \bar{x}_1) + \beta_2(x_2 - \bar{x}_2) + \cdots + \beta_p(x_p - \bar{x}_p)] \\ &= \sum_{i=1}^p (x_i - \bar{x}_i)^2 V[\beta_i] + \sum_{i=1}^{p-1} \sum_{j=1}^p (x_i - \bar{x}_i)(x_j - \bar{x}_j) \text{Cov}[\beta_i, \beta_j] = (\text{式 8}) \end{aligned}$$

第 1 項目は同じ添え字を掛け算する p 個について、第 2 項目は違う添え字同士で共分散を求める $p(p-1)$ 個についてまとめました。ちょっと難しいけど、時間かけて確認ください。いい勉強になります！

(式 8) に $V[\beta_i] = \sigma^2 S^{ii}$, $\text{Cov}[\beta_i, \beta_j] = \sigma^2 S^{ij}$ を代入すると、

$$(\text{式 8}) = \sum_{i=1}^p (x_i - \bar{x}_i)^2 \sigma^2 S^{ii} + \sum_{i=1}^{p-1} \sum_{j=1}^p (x_i - \bar{x}_i)(x_j - \bar{x}_j) \sigma^2 S^{ij} = (\text{式 9})$$

一旦、(式 9) の σ^2 は外した残りを計算しましょう。(式 9') とします。

(式 9') は行列表記できます。

$$\begin{aligned} \bullet (x_i - \bar{x}_i)^2 S^{ii} &= (x_i - \bar{x}_i) S^{ii} (x_i - \bar{x}_i) \\ \bullet (x_i - \bar{x}_i)(x_j - \bar{x}_j) S^{ij} &= (x_i - \bar{x}_i) S^{ij} (x_j - \bar{x}_j) \end{aligned}$$

という形を意識すると、

(式9)=

$$\begin{pmatrix} x_1 - \bar{x}_1 & x_2 - \bar{x}_2 & \dots & x_p - \bar{x}_p \\ S^{11} & S^{12} & \dots & S^{1p} \\ S^{21} & S^{22} & \dots & S^{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S^{p1} & S^{p2} & \dots & S^{pp} \end{pmatrix} \begin{pmatrix} x_1 - \bar{x}_1 \\ x_2 - \bar{x}_2 \\ \vdots \\ x_p - \bar{x}_p \end{pmatrix}$$

とまとめることができ、この式こそ、マハラビノス距離 D^2 と表記できます。

$$(式9)=(式9') \times \sigma^2 = \sigma^2 D^2$$

となります。よって

$$(式7)=V[\beta_1(x_1 - \bar{x}_1) + \beta_2(x_2 - \bar{x}_2) + \dots + \beta_p(x_p - \bar{x}_p)]$$

= (式8,9)と計算して

$$= \sigma^2 D^2 \text{ (マハラビノス距離)}$$

となります。

★V[y]をまとめると

$$V[y] = (式6) = \sigma^2 \left(1 + \frac{1}{n} + D^2\right)$$

とまとめることができます。

(4) 公式と比べるとちょっと違う？

結果を比較すると、

$$\bullet V[y] = (式6) = \sigma^2 \left(1 + \frac{1}{n} + D^2\right) \text{ (式11)}$$

ですが、元々

$$\bullet \sqrt{\left(1 + \frac{1}{n} + \frac{D^2}{n-1}\right)} V_e \text{ (式12)}$$

でしたよね。よく見ると、(式12)を2乗して、 $\sigma^2 = V_e$ として比較すると、

$$\bullet V[y] = \left(1 + \frac{1}{n} + D^2\right) V_e \text{ (式11A)}$$

ですが、元々

$$\bullet V[y] = \left(1 + \frac{1}{n} + \frac{D^2}{n-1}\right) V_e \text{ (式12A)}$$

とちかづきますが、

上の式は D^2 のに、下の式は $\frac{D^2}{n-1}$ と違う？何で？となりますね。

実は、

下の式のDは上の式のDの値に $n-1$ をかけたものとして定義しています。

なので、一緒です。おそらく、推定区間の式は $\frac{V_e}{n_e}$ とnで割ったように見せたいからと思われます。ちゃんと導出すると、別に要らないとわかりました。

QC プラネッツは別に、 $V[y] = \left(1 + \frac{1}{n} + D^2\right) V_e$ としてもOKです。

と、細かくみてきましたが、ちゃんと、 $\sqrt{\left(1 + \frac{1}{n} + \frac{D^2}{n-1}\right)} V_e$ が導出できました。

目的変数 y が

$$y = \bar{y} + \beta_1(x_1 - \bar{x}_1) + \beta_2(x_2 - \bar{x}_2) + \cdots + \beta_p(x_p - \bar{x}_p) + \varepsilon$$

の信頼度 $(100 - \alpha)\%$ の推定区間は、

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \pm t(n - p - 1, \alpha) \sqrt{\left(1 + \frac{1}{n} + \frac{D^2}{n - 1}\right) V_e}$$

で与えられる。(ただし、 D はマハラビノス距離)

の導出がちゃんとできました！ めでたしめでたし！

以上、「重回帰分析の推定区間の式が導出できる」を解説しました。

偏相関係数が導出できる

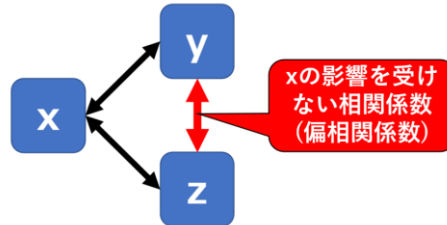
【1】偏相関係数とは

(1) 偏相関係数とは

変数 x と変数 y はある程度の相関関係があり、変数 x と変数 z にもある程度の相関関係があるとします。その場合、変数 y と変数 z の間にも相関関係がありますが、**変数 x の影響が含まれています。**

変数 x の影響を含まない、変数 y と変数 z の相関関係を偏相関係数といいます。

図で描くと、下図のイメージです。



(2) 偏相関係数の式

偏相関係数の式は下の式になります。

$$\rho_{yz,x} = \frac{\rho_{yz} - \rho_{xy}\rho_{xz}}{\sqrt{1 - \rho_{xy}^2}\sqrt{1 - \rho_{xz}^2}}$$

複雑な式ですが、導出過程を理解して偏相関係数の式を理解しましょう。

【2】偏相関係数が導出できる

(1) 偏相関係数の式の理解する方法

導出過程は以下のとおりです。流れで理解しましょう。

1. $y(x), z(x)$ の回帰式を作る
2. 相関係数 ρ 、標準偏差 σ と平方和 S の関係式を駆使して、相関係数 ρ だけで表現した偏相関係数の式を作る
3. $y(x), z(x)$ の残差(x に影響しない誤差の部分)どうしの相関係数が偏相関係数である

では、導出していきます。

(2) 回帰式を作る

y, z について x についての回帰モデルを作ります。その残差を y', z' とします。回帰式を書くと、それぞれ

● $\hat{y} = \bar{y} + a(x - \bar{x})$

● $\hat{z} = \bar{z} + b(x - \bar{x})$

になります。これは単回帰分析の領域なので簡単ですね。なお、 y, z と回帰成分 $\hat{y}, \hat{z}$ と残差成分 y', z' には、

● $y = \hat{y} + y'$

● $z = \hat{z} + z'$

の関係があるので、残差 y', z' は

● $y' = y - (\bar{y} + a(x - \bar{x}))$

● $z' = z - (\bar{z} + b(x - \bar{x}))$

となります。この式を使っていきます。

(2) 関係式をまとめる

① 相関係数 ρ 、標準偏差 σ と平方和 S の関係式

導出できるので、暗記不要です。

相関係数 ρ は2者間つまり単回帰分析の相関係数なので、

$$\rho_{xy} = \frac{s_{xy}}{\sqrt{s_{xx}}\sqrt{s_{yy}}}$$

ですね。これは簡単です。

ここで、平方和の式を展開すると、

$$\bullet S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = n\sigma_x^2$$

$$\bullet S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = n\sigma_y^2$$

(平方和を個数で割ると、分散。分散の平方根は標準偏差の関係を式にしています。すると、

$$\rho_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{S_{xy}}{\sqrt{n\sigma_x^2}\sqrt{n\sigma_y^2}}$$

よって

$$S_{xy} = \rho_{xy} n\sigma_x\sigma_y$$

となります。

② 回帰式の傾き a, b を ρ, σ で表現

また、回帰式の傾き a, b を ρ, σ で表現します。

$$\bullet a = \frac{S_{xy}}{S_{xx}} = \frac{\rho_{xy} n\sigma_x\sigma_y}{n\sigma_x^2} = \rho_{xy} \frac{\sigma_y}{\sigma_x}$$

$$\bullet b = \frac{S_{xz}}{S_{xx}} = \frac{\rho_{xz} n\sigma_x\sigma_z}{n\sigma_x^2} = \rho_{xz} \frac{\sigma_z}{\sigma_x}$$

と表現します。関係式をまとめます。あとで必要なので、

【関係式】

(1) 平方和について

$$\bullet S_{xx} = n\sigma_x^2$$

$$\bullet S_{yy} = n\sigma_y^2$$

$$S_{xy} = \rho_{xy} n\sigma_x\sigma_y$$

(2) 傾きについて

$$\bullet a = \rho_{xy} \frac{\sigma_y}{\sigma_x}$$

$$\bullet b = \rho_{xz} \frac{\sigma_z}{\sigma_x}$$

では、偏相関係数の式を導出します。

(3) 偏相関係数を導出する

偏相関係数を導出する最初の式は

y', z' の相関係数を求める式なので、単純に

$$\rho_{y'z'} = \frac{S_{y'z'}}{\sqrt{S_{y'y'}}\sqrt{S_{z'z'}}}$$

あとは、平方和 $S_{y'y'}, S_{z'z'}, S_{y'z'}$ を ρ だけの式になるように変形していきます。

① 平方和の式を整理

(式 1) において、平方和を展開します。

$$\bullet S_{y'y'} = \sum_{i=1}^n (y'_i - \bar{y}')^2 = \sum_{i=1}^n (y'_i)^2 \quad (\text{ここで、残差の平均 } \bar{y}' \text{ は } 0 \text{ になります})$$

$$= \sum_{i=1}^n ((y - \bar{y}) - a(x - \bar{x}))^2 = \sum_{i=1}^n (y - \bar{y})^2 - 2a \sum_{i=1}^n (y - \bar{y})(x - \bar{x}) + a^2 \sum_{i=1}^n (x - \bar{x})^2$$

よく見ると、それぞれ

$$= S_{yy} - 2as_{xy} + a^2 S_{xx} = n\sigma_y^2 - 2a\rho_{xy} n\sigma_x\sigma_y + a^2 n\sigma_y^2 = n\sigma_y^2 - 2(\rho_{xy} \frac{\sigma_y}{\sigma_x}) \rho_{xy} n\sigma_x\sigma_y + (\rho_{xy} \frac{\sigma_y}{\sigma_x})^2 n\sigma_y^2$$

$$= n\sigma_y^2 (1 - \rho_{xy}^2)$$

となります。

同様に

$$\bullet S_{z'z'} = n\sigma_z^2(1 - \rho_{xz}^2)$$

と $y \rightarrow z$ に変えるだけでOKです。

$$\bullet S_{y'z'} = \sum_{i=1}^n (y'_i - \bar{y}')(z'_i - \bar{z}') = \sum_{i=1}^n (y'_i)(z'_i) = \sum_{i=1}^n ((y - \bar{y}) - a(x - \bar{x}))((z - \bar{z}) - b(x - \bar{x}))$$

(ここで、残差の平均 $\bar{y}', \bar{z}'$ は 0 になります)

$$= \sum_{i=1}^n (y - \bar{y})(z - \bar{z}) - b \sum_{i=1}^n (y - \bar{y})(x - \bar{x}) - a \sum_{i=1}^n (z - \bar{z})(x - \bar{x}) + ab \sum_{i=1}^n (x - \bar{x})^2$$

$$= S_{yz} - bS_{xy} - aS_{xz} + abS_{xx} = \rho_{yz} n\sigma_y\sigma_z - \left(\rho_{xz} \frac{\sigma_z}{\sigma_x}\right) \rho_{xy} n\sigma_x\sigma_y - \left(\rho_{xy} \frac{\sigma_y}{\sigma_x}\right) \rho_{xz} n\sigma_x\sigma_z + \left(\rho_{xy} \frac{\sigma_y}{\sigma_x}\right) \left(\rho_{xz} \frac{\sigma_z}{\sigma_x}\right) n\sigma_x^2$$

$$= n\sigma_y\sigma_z(\rho_{yz} - \rho_{xy}\rho_{xz})$$

平方和を整理すると

$$\bullet S_{y'y'} = n\sigma_y^2(1 - \rho_{xy}^2)$$

$$\bullet S_{z'z'} = n\sigma_z^2(1 - \rho_{xz}^2)$$

$$\bullet S_{y'z'} = n\sigma_y\sigma_z(\rho_{yz} - \rho_{xy}\rho_{xz})$$

② 偏相関係数 $\rho_{y'z'}$ の導出

ここまで来たら、ゴールはもうすぐです。

$$\rho_{y'z'} = \frac{S_{y'z'}}{\sqrt{S_{y'y'}}\sqrt{S_{z'z'}}} = \frac{n\sigma_y\sigma_z(\rho_{yz} - \rho_{xy}\rho_{xz})}{\sqrt{n\sigma_y^2(1 - \rho_{xy}^2)}\sqrt{n\sigma_z^2(1 - \rho_{xz}^2)}} = \frac{\rho_{yz} - \rho_{xy}\rho_{xz}}{\sqrt{1 - \rho_{xy}^2}\sqrt{1 - \rho_{xz}^2}}$$

ちゃんと導出できましたね！

【3】 偏相関係数を計算する

せっかくなので、具体的な偏相関係数の値を求めてみましょう。

(1) データ事例

下のデータを例に挙げます。別にどんなデータでも構いません。

No	年齢x	睡眠時間 y	点数z
1	10	8	56
2	12	7	62
3	14	7.5	64
4	13	9	68
5	10	6.5	72
6	25	6	76
7	22	6.5	80
8	24	7	82
9	22	5.5	84
10	16	7	90

10 人（年齢、睡眠時間ばらばら）であるテストを実施した結果とします。

年齢 x と点数 z、睡眠時間 y と点数 z、年齢 x と睡眠時間 y の関係を考えます。

(2) 相関係数を計算

互いの相関係数を計算しましょう。(計算してみてください)

平方和	値	相関係数	値
Sxx	311.6	Rxy	0.343
Syy	9	pxy	-0.585
Szz	1064.4	Rxz	0.456
Sxy	-31	pxz	0.675
Sxz	388.8	Ryz	0.304
Syz	-54	pyz	-0.552

(3) 偏相関係数を計算

2 者間の相関係数を求める際、もう 1 つの影響を除去した偏相関係数をそれぞれ計算してみましょう。

$$\bullet \rho_{yz,x} = \frac{\rho_{yz} - \rho_{xy}\rho_{xz}}{\sqrt{1-\rho_{xy}^2}\sqrt{1-\rho_{xz}^2}} = -0.262$$

$$\bullet \rho_{xz,y} = \frac{\rho_{xz} - \rho_{xy}\rho_{yz}}{\sqrt{1-\rho_{xy}^2}\sqrt{1-\rho_{yz}^2}} = -0.521$$

$$\bullet \rho_{xy,z} = \frac{\rho_{yz} - \rho_{xz}\rho_{xy}}{\sqrt{1-\rho_{xz}^2}\sqrt{1-\rho_{xy}^2}} = -0.356$$

ただ、偏相関係数がほぼ 0 の場合は、2 者間に相関係数がある程度の値をもっていたとしても相関関係がないことがわかります。2 者間の相関関係を調べる時に活用する偏相関係数がよくわかりましたね。

以上、「偏相関係数が導出できる」を解説しました。

重回帰分析の多重共線性がわかる

【1】多重共線性がわかる

(1) 多重共線性とは

多重共線性を簡単に説明すると、

説明変数と目的変数は正（または負）の相関があるのに、
重回帰モデルを作ると係数が負（または正）と逆の符号がついてしまうこと

例として（あとで事例紹介しますが、）説明変数 x_1, x_2 と目的変数 y があります。説明変数と目的変数には、

● $\rho_{x_1y}=0.772$ （正の相関）

● $\rho_{x_2y}=0.801$ （正の相関）

とあるのに、回帰式を解くと、

$$y = 3.1 - 0.1x_1 + 0.4x_2$$

と負の係数がついてしまう現象のことです。

説明変数 x_1, x_2 も目的変数 y は正の相関があるわけですから、

● x_1 が増加すれば y も増加するはず

● x_2 が増加すれば y も増加するはず

ですが、回帰式は

● x_1 が増加すれば y は減少する式になっています。

(2) 重回帰分析しないと多重共線性はわからない

厄介なことに

多重共線性が起こる条件式は作れない。重回帰分析の結果オーライ的などころがある

なので、解析するしかありません。ただし、よく言われるのが、

目的変数間の相関係数が大きい場合、多重共線性になりやすい

この性質を活用して、多重共線性が起こる場合と起きない場合の事例をこれから解説します。

【2】説明変数が2つの場合の多重共線性

(1) 回帰直線と係数の求め方

説明変数が2つの場合においては、関連記事に説明していますので、ご確認ください。

【関連記事】重回帰分析の回帰式が導出できる

<https://qcplanets.com/method/multi-regression/equation/>

計算結果だけ書くと

● y 切片 β_0 の導出

$$\beta_0 = a = \bar{y} - (b\bar{x}_1 + c\bar{x}_2)$$

から計算し、

●傾き β_1, β_2 の導出

$$S_{11}b + S_{12}c = S_{1y}$$

$$S_{12}b + S_{22}c = S_{2y}$$

を満たす連立方程式から、 β_1, β_2 が導出できます！

つまり、

●傾き β_1, β_2 の導出

$$\beta_1 = \frac{1}{S_{11}S_{22} - S_{12}^2} (S_{22}S_{1y} - S_{12}S_{2y})$$

$$\beta_2 = \frac{1}{S_{11}S_{22} - S_{12}^2} (-S_{12}S_{1y} + S_{11}S_{2y})$$

なお、

● 寄与率 $R = \frac{S_{12}^2}{S_{11}S_{22}} \leq 1$ より $S_{11}S_{22} - S_{12}^2 \geq 0$ なので、
 傾き β_1 、 β_2 の正負は、分子の
 $\beta_1 \rightarrow (S_{22}S_{1y} - S_{12}S_{2y})$
 $\beta_2 \rightarrow (-S_{12}S_{1y} + S_{11}S_{2y})$
 の正負で決まりますね。

(2) 多重共線性が無い事例

下表のデータを用意します。さて、ここで問題です。良問なので是非解いてみてください。

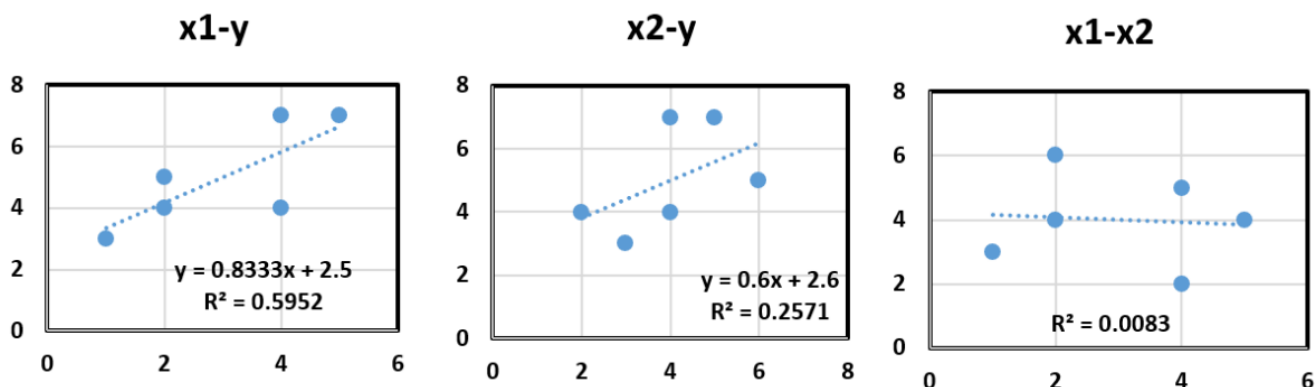
x1	x2	y
1	3	3
4	2	4
2	4	4
5	4	7
4	5	7
2	6	5

【問題】 上表において、回帰直線 $y = a + bx_1 + cx_2$ の定数 a, b, c を求めよ。

結果は

- $y = -0.429 + 0.891x_1 + 0.689x_2$
- 回帰平方和 $S_R = 13.04$
- 残差平方和 $S_e = 0.960$
- 総平方和 $S_T = 14$

- 説明変数 x_1 と目的変数 y との相関関係
 - 説明変数 x_2 と目的変数 y との相関関係
 - 説明変数 x_1 と説明変数 x_2 との相関関係
- をそれぞれ見ると下図のようになります。



たしかに、

- 説明変数 x_1 と目的変数 y とは、正の相関関係
- 説明変数 x_2 と目的変数 y とは、正相関関係
- 説明変数 x_1 と説明変数 x_2 は相関性なし

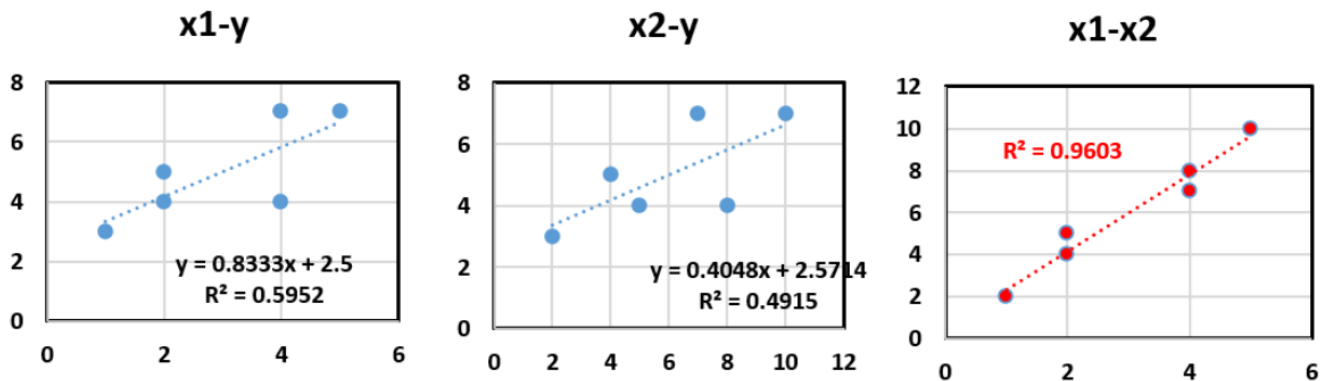
がわかりますね。回帰式の傾きの値も相関関係と同じ正になりました。この場合は多重共線性にならないとなります。

(3) 多重共線性が起きる事例

では、先の表の x_2 を下表のように値をあえて変えてみましょう。

x_1	x_2	y
1	2	3
4	8	4
2	5	4
5	10	7
4	7	7
2	4	5

値の違いを図で確認しましょう。



つまり、説明変数 x_1, x_2 の相関係数を 1 にあえて近づけてみました。

さて、ここで問題です。良問なので是非解いてみてください。

【問題】 上表において、回帰直線 $y = a + bx_1 + cx_2$ の定数 a, b, c を求めよ。

結果は

- $y = 2.9 + 2.3x_1 - 0.8x_2$
- 回帰平方和 $S_R = 8.33$
- 残差平方和 $S_e = 5.67$
- 総平方和 $S_T = 14$

たしかに、

- 説明変数 x_1 と目的変数 y との正の相関関係
 - 説明変数 x_1 と目的変数 y との正の相関関係
 - 説明変数 x_1 と説明変数 x_2 との正の相関関係
- がわかりますね。

ところが、回帰式の傾きの 1 つが逆符号の結果になりました。多重共線性になったとなります。

(4) 多重共線性が起こる理由

- 傾き β_1, β_2 の正負は、分子の
 $\beta_1 \rightarrow (S_{22}S_{1y} - S_{12}S_{2y})$
 $\beta_2 \rightarrow (-S_{12}S_{1y} + S_{11}S_{2y})$
 の正負で決まりますね。

実際に重回帰分析しないと傾きの正負と相関関係の正負の一致・不一致はわからない。ただ、相関性の強い説明変数を用意すると多重共線性が起こることが確認できました。

2変数でも多重共線性が起こることがわかりましたし、手計算できるレベルなので、実際解くと実感がわくはず。次は、教科書によく出て来る、説明変数 3 つの場合について解説します。変数の種類が増えただけで考え方は同じです。

【3】説明変数が 3 つの場合の多重共線性

(1) 多重共線性がない事例

—	x 1	x2	x3	y
1	2	5	4	21
2	0	4	10	15
3	3	10	8	32
4	11	15	26	60
5	14	8	15	33
6	8	8	14	50
sum	38	50	77	211
ave	6.33	8.33	12.83	35.17

互いの相関係数を計算します。

相関係数	ρ_{x1}	x2	x3
x1	1	0.591	0.724
x2	0.591	1	0.805
x3	0.724	0.805	1
y	0.684	0.865	0.828

回帰直線と分散分析を求めます。Excel や python 使いましょう。手計算はキツイです。

—	平方和	自由度	平均平方	F	—	係数
回帰	1183.448	3	394.483	2.865	切片	4.234
残差	275.385	2	137.692	—	X 値 1	0.524
合計	1458.833	5	—	—	X 値 2	2.442
—	—	—	—	—	X 値 3	0.566

たしかに、どの傾きの値も正で、相関係数の正と一致していますね。これは確かに多重共線性が起きていません。

(2) 多重共線性が起きる事例

では、 x_2 のデータを変えてみましょう。変えたことによって相関係数の変化と回帰直線の傾きの値に注目です。

—	x 1	x2	x3	y
1	2	4	4	21
2	0	1	10	15
3	3	5	8	32
4	11	18	26	60
5	14	30	15	33
6	8	15	14	50
sum	38	73	77	211
ave	6.333	12.167	12.833	35.167

マーカー部が変化したところです。相関係数も解きましょう。

相関係数p	x1	x2	x3
x1	1	0.982	0.724
x2	0.982	1	0.616
x3	0.724	0.616	1
y	0.684	0.546	0.828

確かに、マーカ一部分の相関係数は1に上がったのがわかりますね。そうすると、分散分析と回帰直線の傾きを計算すると、下表になります。

—	平方和	自由度	平均平方	F	—	係数
回帰	1351.303	3	450.434	8.378	切片	21.524
残差	107.53	2	53.765	—	X 値 1	13.577
合計	1458.833	5	—	—	X 値 2	-5.831
—	—	—	—	—	X 値 3	-0.109

確かに、正の相関係数であるが、回帰直線の傾きは負になる多重共線性が起こることがわかりますね。

(3) 多重共線性が起こる理由

計算して結果オーライなので、数学的に理由を説明するのは難しいですね。

【4】多重共線性より重回帰分析の結果の妥当性を考えるべき

多重共線性が起こる・起こらないより結果の妥当性やその結果になる原因や因果関係を考察する方がはるかに重要です。

解析方法は手段にすぎません。目的を明確に、結果を考えるようにしましょう。

以上、「重回帰分析の多重共線性がわかる」を解説しました。

【1】単位を変えた場合

(1) 理解に必要な公式と関連記事

本記事で使う重回帰分析の公式とその導出過程を詳細にまとめた関連記事を紹介します。しっかり確認しましょう。なお、本記事では、わかりやすさを優先するために、説明変数が 2 つの場合について解説します。

★おさえておきたい公式

- 平方和の分解: $S_T = S_R + S_e$ (総平方和)=(回帰平方和)+(残差平方和)
- 回帰平方和: $S_R = \beta_1 S_{1y} + \beta_2 S_{2y}$
- 平方和
 - ・ $S_{11} = \sum_{i=1}^n (x_1 - \bar{x}_1)^2$ ・ $S_{22} = \sum_{i=1}^n (x_2 - \bar{x}_2)^2$ ・ $S_{12} = \sum_{i=1}^n (x_1 - \bar{x}_1)(x_2 - \bar{x}_2)$
 - ・ $S_{1y} = \sum_{i=1}^n (x_1 - \bar{x}_1)(y - \bar{y})$ ・ $S_{2y} = \sum_{i=1}^n (x_2 - \bar{x}_2)(y - \bar{y})$
- 傾き β_1, β_2 の導出

$$S_{11}b + S_{12}c = S_{1y} \quad S_{12}b + S_{22}c = S_{2y}$$

$$\beta_1 = \frac{1}{S_{11}S_{22} - S_{12}^2} (S_{22}S_{1y} - S_{12}S_{2y})$$

$$\beta_2 = \frac{1}{S_{11}S_{22} - S_{12}^2} (-S_{12}S_{1y} + S_{11}S_{2y})$$

すべて、計算で解けます。公式暗記は禁物です。解けない場合は関連記事で解けるようにしましょう！

【関連記事】平方和の分解と分散分析ができる(重回帰分析)

<https://qcplanets.com/method/multi-regression/anova-basic/>

(2) 変わるものと変わらないものがある

ここで、説明変数 x_1 の単位が変わって、 $x'_1 = ax_1$ (a は定数倍) に変化したとしましょう。すると、結論は、

- x'_1 の回帰直線の傾き β'_1 が $\frac{1}{a}\beta_1$ に変化する。
- x_2 の回帰直線の傾きは変わらない。
- 総平方和、回帰平方和、残差平方和は変わらない。

(3) 事例

実際に下表データにおいて、説明変数 x_1 を 1/100 に変えて、それぞれ重回帰分析しましょう。下表では、左側から右側を比較しましょう。

x_1	x_2	y	$\Rightarrow$	x'_1	x_2	y
1	3	3	$\Rightarrow$	0.01	3	3
4	2	4	$\Rightarrow$	0.04	2	4
2	4	4	$\Rightarrow$	0.02	4	4
5	4	7	$\Rightarrow$	0.05	4	7
4	5	7	$\Rightarrow$	0.04	5	7
2	6	5	$\Rightarrow$	0.02	6	5

重回帰分析すると下の結果になります。一度、計算して確かめてみてください！いい練習になります！

	x_1 の場合	$x'_1 = \frac{x_1}{100}$ の場合
y 切片 β_0	-0.429	-0.429
x_1 の傾き β_1	0.891	89.076
x_2 の傾き β_2	0.691	0.691
総平方和 S_T	14	14
回帰平方和 S_R	8.333	8.333
残差平方和 S_e	5.667	5.667

確かに、

- x'_1 の回帰直線の傾き β'_1 が $\frac{1}{a}\beta_1$ に変化する。 ($a=100$)
- x_2 の回帰直線の傾きは変わらない。
- 総平方和、回帰平方和、残差平方和は変わらない。

この理由を数式で証明しましょう。数式で理解するとよくわかります!

【2】単位を変えた説明変数の傾きだけが変わる理由

(1) 傾きを導出する式

【1】(1)の公式に対して、説明変数 x_1 が x'_1 に変化するのので、回帰直線の傾きも変化するかもしれません。なので、 $\beta_1 \Rightarrow \beta'_1$ 、 $\beta_2 \Rightarrow \beta'_2$ とします。

説明変数は

- $x'_1 = ax_1$
- $x'_2 = x_2$

ですから、平方和の式に代入しましょう。

(2) 平方和の変化

代入します。

● 平方和

$$\cdot S_{1'1'} = \sum_{i=1}^n (x'_1 - \bar{x}'_1)^2 = \sum_{i=1}^n (ax_1 - a\bar{x}_1)^2 = a^2 \sum_{i=1}^n (x_1 - \bar{x}_1)^2 = a^2 S_{11}$$

ですね。同様に、

$$\cdot S_{2'2'} = S_{22} \quad \cdot S_{1'2'} = aS_{12} \quad \cdot S_{1'y} = aS_{1y} \quad \cdot S_{2'y} = S_{2y}$$

となります。

(3) 単位を変えた説明変数の傾きだけが変わる理由

傾きの式に x'_1, x'_2 を代入します。

$$\begin{aligned} \beta'_1 &= \frac{1}{S_{1'1'} - S_{2'2'} S_{1'2'}^2} (S_{2'2'} S_{1'y} - S_{1'2'} S_{2'y}) & \beta'_2 &= \frac{1}{S_{1'1'} S_{2'2'} - S_{1'2'}^2} (-S_{1'2'} S_{1'y} + S_{1'1'} S_{2'y}) \\ &= \frac{1}{a^2 S_{11} S_{22} - a^2 S_{12}^2} (a S_{22} S_{1y} - a S_{12} S_{2y}) & &= \frac{1}{a^2 S_{11} S_{22} - a^2 S_{12}^2} (-a^2 S_{12} S_{1y} + a^2 S_{11} S_{2y}) \\ &= \frac{1}{a} \frac{1}{S_{11} S_{22} - S_{12}^2} (S_{22} S_{1y} - S_{12} S_{2y}) & &= \frac{1}{S_{11} S_{22} - S_{12}^2} (-S_{12} S_{1y} + S_{11} S_{2y}) \\ &= \frac{1}{a} \beta_1 & &= \beta_2 \end{aligned}$$

つまり、

- x'_1 の回帰直線の傾き β'_1 が $\frac{1}{a}\beta_1$ に変化する。
- x'_2 の回帰直線の傾きは変わらない。

【3】単位を変えても回帰、残差平方和は変化しない

(1) 平方和の分解

平方和の式を書きましょう。関連記事にもあるように、
モデル式 $(y - \bar{y}) = (\hat{y} - \bar{y}) + (y - \hat{y})$ は $(\hat{y}$ は回帰直線上にのる値)

$$\sum_{i=1}^n (y - \bar{y})^2 = \sum_{i=1}^n (\hat{y} - \bar{y})^2 + \sum_{i=1}^n (y - \hat{y})^2 \quad \text{となり、} \quad S_T = S_R + S_e \quad \text{ですね。}$$

(2) 平方和、回帰平方和、残差平方和は変化しない

実は、説明変数 x_i が変化して影響を受けるのは、 $\hat{y}$ がある成分です。

よって、

●総平方和 S_T は変わらない

★回帰平方和 S_R を計算

次に回帰平方和 S_R を計算しましょう。

$$S_R = \beta_1 S_{1y} + \beta_2 S_{2y}$$

ですから、

$$S_R = \beta_1' S_{1y}' + \beta_2' S_{2y}'$$

とすると、

$$= \frac{1}{a} \beta_1 a S_{1y} + \beta_2 S_{2y} = \beta_1 S_{1y} + \beta_2 S_{2y} = S_R$$

となります。よって、

●回帰平方和 S_R は変わらないし

●残差平方和 $S_e = S_T - S_R$ も変わらない

つまり、各平方和の成分は変化しないとわかります。

ちゃんと証明できましたね。結論を再掲すると

● x_1' の回帰直線の傾き β_1' が $\frac{1}{a}\beta_1$ に変化する。

● x_2' の回帰直線の傾きは変わらない。

●総平方和、回帰平方和、残差平方和は変わらない。

となっていますね。

以上、「重回帰分析は単位に影響されない理由がわかる(その1 $\mathbf{x}' = \mathbf{a}\mathbf{x}$ の場合)」を解説しました。

重回帰分析は単位に影響されない理由がわかる(その2 $x' = ax + b$ の場合)

その1は $x' = ax$ ですが、今回、定数値 b を加算すると重回帰分析にどう影響するか？を解説します。

本記事も、(その1)と同様に、わかりやすく説明するために説明変数は2つだけの重回帰分析を使って解説します。

【1】説明変数が定数倍変化した場合

(1) ダミー変数導入の時に本記事が必要

本記事の(その2)は $x' = ax + b$ の場合を解説する理由は、**ダミー変数導入の時に本記事が必要**だからです。ダミー変数を使って重回帰分析する場合、ダミー変数をいくらにすればよいか？気になりますよね。

- ダミー変数：0, 1, 2と1ずつ増やすか？
 - ダミー変数：0, 5, 10とか一定数ずつ増やした方がいいのか？
- などです。

0, 1, 2と1ずつ増やしたり、1, 6, 11と増やしたりと考える場合、両者の関係は「5倍して1足す」関係ですね。つまり、本記事の $x' = ax + b$ の場合がベースとなるので、解説が必要なのです！

(2) 説明変数が定数倍変化した場合

本冊子の【重回帰分析は単位に影響されない理由がわかる(その2 $x' = ax + b$ の場合)】の結果をまとめると、

- x'_1 の回帰直線の傾き β_1' が $\frac{1}{a}\beta_1$ に変化する。
- x_2 の回帰直線の傾きは変わらない。
- 総平方和、回帰平方和、残差平方和は変わらない。

【2】説明変数が定数倍に定数値を加算した場合

(1) 理解に必要な公式と関連記事

本記事で使う重回帰分析の公式とその導出過程を詳細にまとめた関連記事を紹介します。しっかり確認しましょう。なお、本記事では、わかりやすさを優先するために、説明変数が2つの場合について解説します。

★おさえておきたい公式

- 平方和の分解： $S_T = S_R + S_e$ (総平方和)=(回帰平方和)+(残差平方和)
- 回帰平方和： $S_R = \beta_1 S_{1y} + \beta_2 S_{2y}$
- 平方和
 - ・ $S_{11} = \sum_{i=1}^n (x_1 - \bar{x}_1)^2$ ・ $S_{22} = \sum_{i=1}^n (x_2 - \bar{x}_2)^2$ ・ $S_{12} = \sum_{i=1}^n (x_1 - \bar{x}_1)(x_2 - \bar{x}_2)$
 - ・ $S_{1y} = \sum_{i=1}^n (x_1 - \bar{x}_1)(y - \bar{y})$ ・ $S_{2y} = \sum_{i=1}^n (x_2 - \bar{x}_2)(y - \bar{y})$
- 傾き β_1 、 β_2 の導出
$$S_{11}b + S_{12}c = S_{1y} \quad S_{12}b + S_{22}c = S_{2y}$$
$$\beta_1 = \frac{1}{S_{11}S_{22} - S_{12}^2} (S_{22}S_{1y} - S_{12}S_{2y})$$
$$\beta_2 = \frac{1}{S_{11}S_{22} - S_{12}^2} (-S_{12}S_{1y} + S_{11}S_{2y})$$
- y切片 β_0 $\beta_0 = \bar{y} - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2$

すべて、計算で解けます。公式暗記は禁物です。解けない場合は関連記事で解けるようにしましょう！

【関連記事】平方和の分解と分散分析ができる(重回帰分析)

<https://qcplanets.com/method/multi-regression/anova-basic/>

(2) 事例

実際に下表データにおいて、

- (i)説明変数 x_1 の場合
 - (ii)説明変数 x_1 を1/100倍に変えた場合
 - (iii)説明変数 x_1 を1/100倍に変えて、さらに7/100足した場合
- それぞれ重回帰分析しましょう。下表では、左側から右側を比較しましょう。

(i) x_1	(ii) $x'_1 = \frac{x_1}{100}$	(iii) $x'_1 = \frac{x_1}{100} + \frac{7}{100}$	x_2	y
1	0.01	0.08	3	3
4	0.04	0.11	2	4
2	0.02	0.09	4	4
5	0.05	0.12	4	7
4	0.04	0.11	5	7
2	0.02	0.09	6	5

重回帰分析すると下の結果になります。

-	(i) x_1	(ii) $x'_1 = \frac{x_1}{100}$	(iii) $x'_1 = \frac{x_1}{100} + \frac{7}{100}$
y切片 β_0	-0.429	-0.429	-6.664
傾き β_1	0.891	89.076	89.076
傾き β_2	0.689	0.689	0.689
総平方和 S_T	14	14	14
回帰平方和 S_R	8.333	8.333	8.333
残差平方和 S_e	5.667	5.667	5.667

確かに、

- x'_1 の回帰直線の傾き β'_1 が $\frac{1}{a}\beta_1$ に変化する。(a = 1/100)
- x_2 の回帰直線の傾きは変わらない。
- x'_1 の回帰直線の y 切片 β'_0 が $\beta_0 - \frac{b}{a}\beta_0$ に変化する。
- 総平方和、回帰平方和、残差平方和は変わらない。

この理由を数式で証明しましょう。数式で理解するとよくわかります!

【3】回帰直線の値と平方和の値の変化を数式で理解する

(1) 数式を準備する

上の【2】(1) 紹介した式をベースとします。

ここで、説明変数 x_1 が x'_1 に変化するので、回帰直線の傾きも変化するかもしれません。なので、 $\beta_0 \Rightarrow \beta'_0$ 、 $\beta_1 \Rightarrow \beta'_1$ 、 $\beta_2 \Rightarrow \beta'_2$ とします。説明変数は

- $x'_1 = ax_1$
- $x'_2 = x_2$

ですから、平方和の式に代入しましょう。

(2) 平方和の変化

代入します。

● 平方和

$$\cdot S_{1'1'} = \sum_{i=1}^n (x'_1 - \bar{x}'_1)^2 = \sum_{i=1}^n ((ax_1 + b) - (\overline{ax_1} + b))^2 = a^2 \sum_{i=1}^n (x_1 - \bar{x}_1)^2 = a^2 S_{11}$$

ですね。同様に、

$$\cdot S_{2'2'} = S_{22} \quad \cdot S_{1'2'} = aS_{12} \quad \cdot S_{1'y} = aS_{1y} \quad \cdot S_{2'y} = S_{2y}$$

となります。

(3) 単位を変えた説明変数の傾きだけが変わる理由
傾きの式に x'_1, x'_2 を代入します。

$$\begin{aligned}\beta'_1 &= \frac{1}{S_{1'1'}S_{2'2'} - S_{1'2'}^2} (S_{2'2'}S_{1'y} - S_{1'2'}S_{2'y}) & \beta'_2 &= \frac{1}{S_{1'1'}S_{2'2'} - S_{1'2'}^2} (-S_{1'2'}S_{1'y} + S_{1'1'}S_{2'y}) \\ &= \frac{1}{a^2S_{11}S_{22} - a^2S_{12}^2} (aS_{22}S_{1y} - aS_{12}S_{2y}) & &= \frac{1}{a^2S_{11}S_{22} - a^2S_{12}^2} (-a^2S_{12}S_{1y} + a^2S_{11}S_{2y}) \\ &= \frac{1}{a} \frac{1}{S_{11}S_{22} - S_{12}^2} (S_{22}S_{1y} - S_{12}S_{2y}) & &= \frac{1}{S_{11}S_{22} - S_{12}^2} (-S_{12}S_{1y} + S_{11}S_{2y}) \\ &= \frac{1}{a} \beta_1 & &= \beta_2\end{aligned}$$

回帰直線 y 切片 β'_0

$$\beta'_0 = \bar{y} - \beta'_1\bar{x}'_1 - \beta'_2\bar{x}'_2 = \bar{y} - \frac{1}{a}\beta_1(a\bar{x}_1 + b) - \beta_2\bar{x}_2 = (\bar{y} - \beta_1\bar{x}_1 - \beta_2\bar{x}_2) - \frac{b}{a}\beta_1 = \beta_0 - \frac{b}{a}\beta_1$$

となりますね。次に平方和を計算してみましょう。

【4】回帰、残差平方和は変化しない

(1) 平方和、回帰平方和、残差平方和は変化しない

実は、説明変数 x_1 が変化して影響を受けるのは、 $\hat{y}$ がある成分です。よって、

- 回帰平方和 S_R は変わらないし
 - 残差平方和 $S_e = S_T - S_R$ も変わらない
- つまり、各平方和の成分は変化しないとわかります。

以上、「重回帰分析は単位に影響されない理由がわかる(その2 $x' = ax + b$ の場合)」を解説しました。

【1】 偏回帰係数に関する検定と推定の式を導出

(1) 偏回帰係数の期待値と分散を導出

偏回帰係数の期待値と分散が本記事のキーポイントとなりますが、結構難しいです。関連記事で詳細に解説していますので、確認しましょう。

本冊子【重回帰分析の推定区間の式が導出できる】で確認しましょう。

★偏回帰係数の期待値と分散

- $E[\beta_j] = \beta_j$
- $V[\beta_j] = \sigma^2 S^{jj}$

(2) 偏回帰係数に関する検定の式を導出

t 分布（正規分布でもいいと思いますけど）で検定するために、 $z = \frac{x - \bar{x}}{\sigma}$ な式を作ります。

(分子)は偏回帰係数に関する検定なので、

- (分子) = $\beta_i - \beta_{i0}$ (β_{i0} は検定で使う値)

分母は分散を使うので、

- $V[\beta_j] = \sigma^2 S^{jj}$

ですが、 σ^2 は未知なので代わりに V_e を使って

- $V[\beta_j] = V_e S^{jj}$

を使うと、検定統計量(t 分布)は $t = \frac{\beta_i - \beta_{i0}}{\sqrt{V_e S^{jj}}}$ を使って、検定します。

- 検定統計量 $t = \frac{\beta_i - \beta_{i0}}{\sqrt{V_e S^{jj}}}$

(3) 偏回帰係数に関する推定の式を導出

検定統計量の式が出来たらあとは、t 分布の推定区間を求める式を作ればいいので、

$$\beta \pm t(\Phi, \alpha) \sqrt{V_e S^{jj}}$$

の式となりますよね。

- 推定区間の式 $\beta \pm t(\Phi, \alpha) \sqrt{V_e S^{jj}}$

【2】 検定・推定を解くための例題

(1) データを用意

下表のような説明変数が 2 つのデータを用意します。

x1	x2	y
8	3	3
11	2	4
9	4	4
12	4	7
11	5	7
9	6	5

(2) 重回帰分析しよう！

●問題です。

上の表を重回帰分析して、下表の各値を求めよ。

変数	値	変数	値
$S_{11} =$	??	$S^{11} =$	??
$S_{12} =$	??	$S^{12} =$	??
$S_{22} =$	??	$S^{22} =$	??
$S_{1y} =$	??	$\beta_0 =$	??
$S_{2y} =$	??	$\beta_1 =$	??
$S_{yy} =$	??	$\beta_2 =$	??
$S_R =$	??	$V_R =$	??
$S_e =$	??	$V_e =$	??
$S_T =$	??	-	-

さっと計算できますか？ S^{ij} は S_{ij} の逆行列の各成分です。

(3) 重回帰分析の結果

答えは下表です。

変数	値	変数	値
$S_{11} =$	12	$S^{11} =$	$0.084(\frac{10}{119})$
$S_{12} =$	-1	$S^{12} =$	$0.0084(\frac{1}{119})$
$S_{22} =$	10	$S^{22} =$	$0.1008(\frac{11}{119})$
$S_{1y} =$	10	$\beta_0 =$	-6.664
$S_{2y} =$	6	$\beta_1 =$	0.891
$S_{yy} =$	14	$\beta_2 =$	0.689
$S_R =$	8.33	$V_R =$	4.165
$S_e =$	5.67	$V_e =$	1.89
$S_T =$	14	-	-

さて、検定と推定をやってみましょう。

【3】 偏回帰係数に関する検定の例題

(1) 例題

偏回帰係数が2つあるので、それぞれ検定しましょう。

問1 偏回帰係数 β_1 が1から変化したかどうかを有意水準 $\alpha=5\%$ で検定せよ。

問2 偏回帰係数 β_2 が0から変化したかどうかを有意水準 $\alpha=5\%$ で検定せよ。

(2) 解法

公式を再掲します。

●検定統計量 $t = \frac{\beta_i - \beta_{i0}}{\sqrt{V_e S^{ii}}}$

① 問1

検定統計量： $t_1 = \frac{\beta_1 - \beta_{10}}{\sqrt{V_e S^{11}}} = \frac{0.891 - 1}{\sqrt{1.89 \times 0.084}} = -0.274$

棄却域(両側検定とします)： $t(\Phi_e, \alpha) = t(3, 0.05) = 3.182$

$|t_1| = 0.274 < 3.182$

より、仮説は棄却されず、「偏回帰係数 β_1 が1から変化したとはいえない。」となります。

②問 2

$$\text{検定統計量} : t_2 = \frac{\beta_2 - \beta_{20}}{\sqrt{V_e S^{22}}} = \frac{0.689 - 0}{\sqrt{1.89 \times 0.1008}} = 1.580$$

$$\text{棄却域(両側検定とします)} : t(\Phi_e, \alpha) = t(3, 0.05) = 3.182$$

$$|t_2| = 1.580 < 3.182$$

より、仮説は棄却されず、「偏回帰係数 β_2 が1から変化したとはいえない。」となります。

【4】偏回帰係数に関する推定の例題

(1) 例題

偏回帰係数が2つあるので、それぞれ検定しましょう。

問1 偏回帰係数 β_1 の95%の区間推定を求めよ。

問2 偏回帰係数 β_2 の95%の区間推定を求めよ。

(2) 解法

公式を再掲します。

$$\bullet \text{推定区間の式 } \beta \pm t(\Phi, \alpha) \sqrt{V_e S^{jj}}$$

$$\text{①問1 : 推定区間 } \beta_1 \pm t(\Phi_e (= 3), \alpha (= 0.05)) \sqrt{V_e S^{11}} = 0.891 \pm 3.182 \times \sqrt{1.89 \times 0.084} = 0.891 \pm 1.268$$

$$\text{②問2 : 推定区間 } \beta_2 \pm t(\Phi_e (= 3), \alpha (= 0.05)) \sqrt{V_e S^{22}} = 0.689 \pm 3.182 \times \sqrt{1.89 \times 0.1008} = 0.689 \pm 1.389$$

できましたね。単なる公式の代入になりがちですが、関連記事を活用して公式の導出も理解しましょう。

以上、「偏回帰係数に関する検定と推定がよくわかる」を解説しました。

【1】変数増減法とは

複数の独立変数の候補の中から、目的変数 y をよく説明する組み合わせを決定する方法です。

【2】最適な変数の種類の見つけ方

(1) 機械的に見つかるものではない

残念ながら、機械的に最適な変数の種類が決まる方法は存在しません。
なぜなら、最適条件を決める方法が目安でしかないから

変数増減法を使って、変数を絞り込めるが、最後は、あなたが変数をそれに決めた理由を説明することが最も大事です。

最適条件を決める方法が目安でしかありませんが、それを知っておくことは重要です。

(2) 変数の数を決める目安

変数を選ぶのが妥当かどうかを決める目安は、重回帰分析で出て来る値の妥当性となります。具体的には、

1. 回帰平方和 S_R と総平方和 S_T の比
2. 回帰の分散分析による F 検定
3. 多重共線性が起きていないかどうかチェック
4. その変数の偏回帰係数の検定

くらいです。結構使えそうな判定条件ですが、説明力に欠ける点があります。

1. 平方和の比の大小を見ても、判定基準はどこにもない(自分で決めるしかない)
2. F 検定、偏回帰係数の検定などの検定結果は絶対正しいという意味ではない
3. 多重共線性の有無は人それぞれの見解になる

つまり、

変数増減法や重回帰分析からいろいろな数字を出してくれるが、判断は自分で決めるしかありません。

変数増減法の手法を暗記するより、むしろ目的とする目的変数 y を表現する説明変数が妥当かどうかは考える方が大事です。

(3) 最後は自分で論理を組むしかない

自分で論理を組んで、妥当性を評価するには、データの特性なり、あなたのデータを評価する人と妥当性を確認するなどがが必要です。数学を使って、機械的に誰もが同じ結果になるわけではありません。変数増減法という手段を解くことが目的化しないよう注意が必要です。

【3】変数増減法の例題

最後は、自分で考えて結論を出す変数増減法ですが、具体事例を見ながら変数の種類による変化を見ていきましょう。

(1) データ事例

説明変数が最大 3 種類あるデータを用意して、変数を増やしながら重回帰分析の変化を見て来ましょう。使うデータを下表のようにします。

No	x1	x2	x3	y
1	10	8	4	56
2	12	7	5	62
3	14	7.5	3	64
4	13	9	11	68
5	10	6.5	8	72
6	25	6	5	76
7	22	6.5	6	80
8	25	7	7	82
9	23	5.5	10	80
10	16	7	21	90
sum	170	70	80	730
ave	17	7	8	73

(2) 説明変数が 1 つの場合

まず、説明変数 **x1** だけの場合の単回帰分析をやってみましょう。

No	x1	x2	x3	y
1	10	-	-	56
2	12	-	-	62
3	14	-	-	64
4	13	-	-	68
5	10	-	-	72
6	25	-	-	76
7	22	-	-	80
8	25	-	-	82
9	23	-	-	80
10	16	-	-	90
sum	170	-	-	730
ave	17	-	-	73

単回帰分析の結果をまとめます。一度は手を動かして計算してみてくださいね。いい練習になります。

分散分析	平方和 S	自由度 ϕ	平均平方	寄与率 R	回帰直線	
回帰R	436.26	1	436.26	0.662	y切片 β_0	53.68
残差e	557.74	8	69.72	-	傾き β_1	1.136
合計T	994	9	-	-	傾き β_2	-
-	-	-	-	-	傾き β_3	-

(3) 説明変数を 1⇒2 に増やした場合

次に説明変数 **x2** を追加して単回帰分析から重回帰分析に切り替えてやってみましょう。これも手計算で確認してみましょう。いい練習になります。

No	x1	x2	x3	y
1	10	8	–	56
2	12	7	–	62
3	14	7.5	–	64
4	13	9	–	68
5	10	6.5	–	72
6	25	6	–	76
7	22	6.5	–	80
8	25	7	–	82
9	23	5.5	–	80
10	16	7	–	90
sum	170	70	–	730
ave	17	7	–	73

重回帰分析の結果をまとめます。一度は手を動かして計算してみてくださいね。いい練習になります。

分散分析	平方和 S	自由度 ϕ	平均平方	–	寄与率 R	–	回帰直線	
回帰R	457.15	2	228.58	–	0.678	–	y切片 β_0	69.97
残差e	536.85	7	76.69	–	–	–	傾き β_1	0.95
合計T	994	9	–	–	–	–	傾き β_2	-1.88
–	–	–	–	–	–	–	傾き β_3	–

(4) 説明変数を 2⇒3 に増やした場合

次に説明変数 x3 を重回帰分析します。さすがに説明変数が 3 つあると手計算はキツイので、Excel を使しましょう。

No	x1	x2	x3	y
1	10	8	4	56
2	12	7	5	62
3	14	7.5	3	64
4	13	9	11	68
5	10	6.5	8	72
6	25	6	5	76
7	22	6.5	6	80
8	25	7	7	82
9	23	5.5	10	80
10	16	7	21	90
sum	170	70	80	730
ave	17	7	8	73

重回帰分析の結果をまとめます。

分散分析	平方和 S	自由度 ϕ	平均平方	–	寄与率 R	–	回帰直線	
回帰R	905.66	3	301.89	–	0.955	–	y切片 β_0	60.93
残差e	88.34	6	14.72	–	–	–	傾き β_1	0.93
合計T	994	9	–	–	–	–	傾き β_2	-2.07
–	–	–	–	–	–	–	傾き β_3	1.35

と3回、回帰分析したので、違いを比較してみましょう。

【4】変数の数と重回帰分析の影響

(1) 説明変数の種類による回帰分析の違い

説明の種類と平方和・寄与率・回帰直線の違いを下表にまとめました。

平方和	x1	x1,x2	x1,x2,x3
回帰平方和 S_R	436.26	457.15	905.66
残差平方和 S_e	557.74	536.85	88.34
総平方和 S_T	994	994	994
寄与率	0.662	0.678	0.955
回帰直線	x1	x1,x2	x1,x2,x3
y切片 β_0	53.69	69.97	60.93
傾き β_1	1.13	0.95	0.93
傾き β_2	-	-1.88	-2.07
傾き β_3	-	-	1.35

(2) 説明変数を増やすと回帰平方和 S_R と寄与率 R は増大

基本的には、説明変数を増やすと回帰平方和 S_R と寄与率 R は増大し

- 回帰平方和 S_R は総平方和 S_T に近づく
- 寄与率 R は1に近づく

多重共線性によって、正の相関があるのに、回帰直線の傾きが負になることがありますが、基本的には、説明変数を増やすと回帰平方和 S_R と寄与率 R は増大します。

式で書くと、

● $S_R = \frac{S_{xy}^2}{S_{xx}}$ (単回帰の場合)

● $S_R = \beta_1 S_{1y} + \beta_2 S_{2y}$ (重回帰2変数の場合)

● $S_R = \beta_1 S_{1y} + \beta_2 S_{2y} + \beta_3 S_{3y}$ (重回帰3変数の場合)

...

となっていくます。

重回帰分析の S_R は加算する項が説明変数の数だけ増えるし、平方和 S_{ir} はすべて正なので、偏回帰係数 β_i がすべて正なら、回帰平方和 S_R は単調増加し総平方和 S_T に近づきますよね！

なので、寄与率も増加し精度の高い回帰直線ができるわけですが、それだけよいかどうかいろいろ疑問に思いますよね。だから、いろいろ検定したりチェックして説明変数の種類の妥当性を確認する必要があります。説明変数をいろいろいじって重回帰分析しても

1. 平方和の比の大小を見ても、判定基準はどこにもない(自分で決めるしかない)
2. F検定、偏回帰係数の検定などの検定結果は絶対正しいという意味ではない
3. 多重共線性の有無は人それぞれの見解になる

なので、最終的にはあなたの論理で重回帰分析に用いる説明変数を決めることになります。

以上、「変数増減法がよくわかる」を解説しました。