



統計学

最小二乗法 7 多重共線性

担当： 長倉 大輔
(ながくらだいすけ)

多重共線性

■ 完全な多重共線性と弱い多重共線性

線形回帰モデルの分析において、多重共線性という問題がある。多重共線性には「**完全な多重共線性**」と呼ばれるものと「**弱い多重共線性**」と呼ばれるものがあり、これら2つは問題の性格が実はかなり違う。

完全な多重共線性は深刻な問題を起こすが、それを回避するのは容易であり、起きているかどうかもすぐわかる。

弱い多重共線性は、まず起きているかどうかの判断が明確にはできなく、またこれは観測数が十分ではないという問題と本質的には同じなので起きたとしても分析者は基本的にどうすることもできない。

多重共線性

- 完全な多重共線性

まず完全な多重共線性についてみていく。

以下にみるように完全な多重共線性が起こると以下の問題が起こる。

「係数の推定ができなくなる。」

これに付随して、もちろん t 検定もできなくなるし、係数の推定を前提としたすべての分析ができなくなる。

多重共線性

以下の重回帰モデルを考えよう。

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \dots + \beta_K X_{Ki} + \varepsilon_i$$

全ての i について $X_{0i} = 1$ であると考えることができるので、このモデルは定数項のあるモデルも特別な場合として含んでいることに注意。この時、説明変数 $X_{0i}, X_{1i}, \dots, X_{Ki}$ が線形従属している、すなわち

$$\sum_{k=0}^K \delta_k X_{ki} = 0$$

が全ての i に対して、少なくとも1つが0でない $\delta_k, k = 0, \dots, K$ に対して成り立っているとき、**完全な多重共線性**が起こっているという。

多重共線性

以下では

- 完全な多重共線性が起こっているときは係数を推定することができない。

ことを例を通じて確認する。

例: 説明変数が X_{0i} と X_{1i} の 2 つで $X_{0i} = 2X_{1i}$ が成り立つとする。この時 $\delta_0 = -1$ と $\delta_1 = 2$ に対して

$$\sum_{k=0}^1 \delta_k X_{ki} = 0$$

が成り立つので完全な多重共線性が起こっている。

多重共線性

この時

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \varepsilon_i$$

の β_0 と β_1 を最小二乗法で推定する問題を考えてみよう。

残差平方和 $SSR(b_0, b_1)$ を微分して0とおいた式は

$$\sum_{i=1}^N (Y_i - b_0 X_{0i} - b_1 X_{1i}) X_{0i} = 0$$

および

$$\sum_{i=1}^N (Y_i - b_0 X_{0i} - b_1 X_{1i}) X_{1i} = 0$$

である。

多重共線性

ここで $X_{0i} = 2X_{1i}$ を代入すると、これら2つの式はそれぞれ

$$\begin{aligned} 2\sum_{i=1}^N (Y_i - 2b_0 X_{1i} - b_1 X_{1i}) X_{1i} &= 0 \\ \Rightarrow \sum_{i=1}^N (Y_i - 2b_0 X_{1i} - b_1 X_{1i}) X_{1i} &= 0 \end{aligned}$$

および

$$\sum_{i=1}^N (Y_i - 2b_0 X_{1i} - b_1 X_{1i}) X_{1i} = 0$$

となる。つまりこれら2つの式は結局

$$\sum_{i=1}^N Y_i X_{1i} - (2b_0 + b_1) \sum_{i=1}^N X_{1i}^2 = 0$$

という同じ式が2つあるのと同じである。

多重共線性

この式は

$$2b_0 + b_1 = \frac{\sum_{i=1}^N Y_i X_{1i}}{\sum_{i=1}^N X_{1i}^2}$$

と書ける。右辺はデータが与えられれば定数である。

この式を満たす b_0 と b_1 の値は無数にある。よって残差平方和を最小にする b_0 と b_1 は無数にあり、1つに決めることができない

→ β_0 と β_1 は推定できない、
となる。

多重共線性

より直観的には

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \varepsilon_i$$

という式において $X_{0i} = 2 X_{1i}$ が成り立つので、これを代入すると、この式は

$$Y_i = (2\beta_0 + \beta_1) X_{1i} + \varepsilon_i = \gamma X_{1i} + \varepsilon_i$$

と表せる。よって $\gamma = 2b_0 + b_1$ を満たすどのような b_0 と b_1 の組み合わせに対してすべて

$$Y_i = (2b_0 + b_1) X_{1i} + \varepsilon_i = \gamma X_{1i} + \varepsilon_i$$

という同じ関係式が成り立つので、この β_0, β_1 と b_0 と b_1 は観測上全く区別がつかないだろうことがわかる。

多重共線性

例題1: 先ほどのモデル

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \varepsilon_i$$

において、完全なる多重共線性が成り立っているとき、 X_{0i} と X_{1i} の相関係数の絶対値は 1 になることを確かめなさい。

多重共線性

■ 完全な多重共線性への対処法

完全な多重共線性が起こった場合、最小二乗法による推定ができなくなるので、何らかの統計ソフトで最小二乗法によって重回帰モデルを推定すると、通常エラーがおこるからすぐにわかる(ただエクセルで推定した場合、出力自体は普通に出てきた。ただし標準誤差の値が0だったりして結果がおかしかった)。

その場合対処法としては、線形従属している変数の内、1つを除いてあげればよい。

多重共線性

■ 完全な多重共線性の例(ダミー変数)

ダミー変数のところに出てきた、男性ダミーを入れた賃金方程式

$$Y_i = \alpha + \alpha_M D_{M,i} + \beta X_i + \varepsilon_i, \\ i = 1, \dots, n, \varepsilon_i \sim i.i.d. N(0, \sigma^2)$$

Y_i : 賃金の対数値,

X_i : 就業年数,

$D_{M,i}$: 男性ダミー.

を考えよう。

多重共線性

ここで、この式に女性ダミー $D_{F,i}$ も入れるとどうなるだろうか？

女性ダミーも入れた式は

$$Y_i = \alpha + \alpha_M D_{M,i} + \alpha_F D_{F,i} + \beta X_i + \varepsilon_i ,$$

となる。

この時、この重回帰式には**完全な多重共線性が発生している**。これを見てみよう。

多重共線性

定数項 α は説明変数 $X_{0,i} = 1$ (すべての i) と係数 α を掛け合わせたものとみなせることに注意。

この時、説明変数の線形結合

$$\delta_0 + \delta_1 D_{M,i} + \delta_2 D_{F,i} + \delta_3 X_i$$

を考えると、例えば、 $\delta_0 = 1, \delta_1 = -1, \delta_2 = -1, \delta_3 = 0$ の時に (すべての i について)

$$\delta_0 + \delta_1 D_{M,i} + \delta_2 D_{F,i} + \delta_3 X_i = 0$$

が成り立つので多重共線性が発生していることがわかる。よってどちらかのダミー変数を抜くか、定数項を除く必要がある。

多重共線性

一般に(定数項を含んだモデルにおいては)、回帰式に J 個のダミー変数, $D_{j,i}, j=1,\dots,J$ を含めた場合、すべての i について

$$\sum_{j=1}^J D_{j,i} = 1$$

が成り立つと完全な多重共線性が発生する。この場合、どれか一つのダミー変数を除くか、定数項を除く必要がある。

このように(複数の)ダミー変数を用いて分析をする場合、多重共線性の問題に注意する必要がある。

多重共線性

例題2

パネルデータ分析に出てきた固定効果モデルについて、
個体間で値は異なるが、時間を通じて一定の説明変数
を入れてしまうと多重共線性の問題が起こることを確認
しなさい。

多重共線性

■ 弱い(意味での)多重共線性

次に**弱い多重共線性**についてみていく。

弱い多重共線性は以下のような現象である。

先ほどの重回帰モデルに対して、説明変数 $X_{0i}, X_{1i}, \dots, X_{Ki}$ の線形結合の値が 0 に近い場合、すなわち

$$\sum_{k=0}^K \delta_k X_{ki} \approx 0$$

が全ての i に対して成り立っているとき、**弱い多重共線性**が起こっているという。

多重共線性

弱い多重共線性は定義が若干あいまいである。

例えばどれくらい 0 に近ければ弱い多重共線性といえるのか、明確な基準はない。

弱い多重共線性が起こっても 係数 $\beta_0, \dots, \beta_K$ の推定は最小二乗法で問題なくできる。

では弱い多重共線性は何が問題なのだろうか？

以下のような “問題” が起こる。

多重共線性

■ 係数の標準誤差が大きくなる。

これは $\sum_{k=0}^K \delta_k X_{ki}$ の値が 0 に近ければ近いほど大きくなる。

この問題に起因してどのようなことが起こるかというと

- ① 弱い多重共線性が起こっている(さらに δ_k が0でない)
 X_{ki} を全て回帰式に入れるとそれらの t 値が有意にならないが、いずれかを抜くと(場合によっては推定値が大きく変わり)、有意になる。
- ② 観測数を増やすと推定値が大きく変動する場合がある(推定が不安定)。これは標準誤差が大きいことに起因。

多重共線性

■ 弱い多重共線性への対処法

対処法: **特**にない。

まず最初に述べたように弱い多重共線性が起きているかどうか(線形結合がどれだけ0に近ければ弱い多重共線性が起きていると言えるのか)について目安がない(他にも多重共線性を諮る尺度はいくつか提案されているが、一致した見解は得られていない)。

1つの方法としては、多重共線性を引き起こしていそうな(かつ分析にあまり重要でない変数)を回帰式から除くという方法があるが、恣意的であり、根本的な解決にはならない。

多重共線性

また弱い多重共線性は、標準誤差、つまり推定量の分散が大きくなってしまう事による問題なので、(完全な多重共線性でなければ)観測数が十分大きければ原理的には解決する(標本が大きくなるにつれて分散は小さくなる)。

つまり弱い多重共線性の問題は実は**観測数が少ない**という問題と本質的には同じである。

さらに、弱い多重共線性が起きていても、統計的な分析としては実は全く問題ない(標準誤差は大きいがそれは正確に推定でき、よって**推定精度がよくないことが正しくわかる**)ので、そもそも弱い多重共線性が問題かどうかというと人によっては問題ないという人もいる。

多重共線性

多重共線性の問題をまとめると以下のようになる

(完全な多重共線性)

原因: 説明変数が線形従属している。

症状: 最小二乗法によって係数の推定ができなくなる。

対処法: 線形従属している説明変数のどれかを抜く。

深刻度: 症状は深刻だがすぐわかり、対処が容易なのでそれほど深刻な問題ではない。

(弱い多重共線性)

原因: 説明変数が線形結合が0に近い。

症状: 最小二乗法による推定量の標準誤差が大きくなる。

対処法: 特にない。

深刻度: 起きているかどうかの判断が恣意的であり、また起きていたとしても統計分析的には問題ない。^{o22}

演習問題

問題1 次のダミー変数のうち、定数項のある回帰式に
いられない組み合わせどのようなになるか？

- ① 男性なら1, 女性なら 0 を取るダミー変数
- ② 女性なら1, 男性なら 0 を取るダミー変数
- ③ 30歳以上なら1、30歳未満なら 0 を取るダミー変数
- ④ 30歳未満なら1、30歳以上なら 0 を取るダミー変数
- ⑤ 所得が100万円未満なら1、100万円以上なら0 を取るダミー変数
- ⑥ 所得が100万円以上500万円未満なら 1、それ以外なら 0 をとるダミー変数
- ⑦ 所得が500万円以上なら 1、未満なら 0 を取るダミー変数
- ⑧ 20歳以上なら 1、未満なら 0 を取るダミー変数。
- ⑨ 所得が1000万円以上なら 1、未満なら0 を取るダミー変数

演習問題

問題2. 回帰式 $Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \varepsilon_i$ を推定したいとする。ここで説明変数を見てみると、すべての i について $X_{0i} = 0.5 X_{1i}$ が成り立っていることが分かった。

(1) これは完全な多重共線性か弱い多重共線性かどちらの問題か？

(2) X_{0i} を除いて、 $Y_i = \gamma X_{1i} + \varepsilon_i$ の γ を最小二乗法で推定した。この時 γ は観測数が増えるにつれて、どのような値(真の値)に近づいていくか？

(3) X_{1i} を除いて $Y_i = \alpha X_{0i} + \varepsilon_i$ の α を最小二乗法で推定した。この時 α は観測数が増えるにつれて、どのような値(真の値)に近づいていくか？