回帰分析と確率変数

九州大学　鈴木譲

1 目的

この報告では、回帰分析における確率変数の扱いについて考察する。回帰分析においては、標本の各要素に関して、従属変数の実測値をどのように扱うかによってアプローチが分かれる。この値を定数として扱うのが決定論的モデルであり、確率変数の取る値として扱うのが確率モデルである。

回帰分析の理論に特化した文献では、記述統計と推測統計の区別をせずに一律に確率モデルを用いて議論を展開しているものが多い。本報告では、このような方法が社会学において適切であるかどうかを吟味する。

2 方法

まず、決定論的モデルと確率モデルとの相異を吟味する。多くの場合、この 2 つのモデルの相異は、回帰分析における従属変数を非確率変数、確率変数のいずれの観点からとらえるかであるとされる。しかしながら、この説明は厳密には正しくないことを示す。

次に、確率モデルを記述統計、推測統計に一律に適用する方法の利点と問題点について述べる。この点に関連して、従属変数が二値変数の場合に、従属変数の値を直接用いて回帰分析を行う問題点として、不均一分散性（heteroskedasticity）がしばしばあげられるが、この議論が不十分であることを示す。また、ロジスティック回帰分析では、二値変数の発生確率を従属変数とするが、確率の根拠を吟味しなくてはならないことを示す。

3 結果

記述統計と推測統計を区別せずに、一律に確率モデルを適用する方法には大きく 2 つの問題があり、社会学においては適切とは考えられない。第一の問題は、多くの場合に確率の根拠が脆弱である点であり、第二の問題は記述統計と推測統計とで、確率の根拠が異なってしまう点である。

ロジスティック回帰分析においては、標本抽出を前提としているわけではないから、確率の根拠は標本抽出ではない。従って安易にこの手法を用いると、確率の根拠が不明確であるにもかかわらず二値変数の発生確率を従属変数として設定することになり、結局は確率モデルを一律に適用する場合と同様の誤謬を犯すことになる。

4 結論

社会学において回帰分析を用いる際には、記述統計では決定論的モデルを用い、推測統計では確率モデルを用い、両者を分けて考えるべきである。また、ロジスティック回帰分析を適用する場合には、従属変数の確率の根拠を吟味することが不可欠である。すなわち、二値変数の取る値が定数ではなく、確率変数の取る値としてとらえることができるかを吟味しなくてはならない。言い換えれば、標本の各要素に関して、二値変数に対応する現象が、確率的に発生すると考えられるかどうかである。

なお、ロジスティック回帰分析においては、3 種類の確率が用いられるのでこれらを区別する必要がある。1 つ目は、二値変数の発生確率であり、これは離散確率である。2 つ目は、ロジスティック分布であり、これは一般化線型モデル（Generalized Linear Model）のリンク関数の逆関数である。3 つ目は、標本抽出に伴う標本分布である。