比率の変動要因を分析する際の注意事項

実証分析においてなんらかの比率の動きについて分析したい事がある。例えば、人口比や所得比、等々がどのような要因によって決定されるかなどを分析する時である。線形回帰分析を用いてこのような比率を分析する際には注意が必要である。例えば、ある変数 X のある変数 Y に対する比率 R は、$R = X/Y$ で定義される。X も Y も正の値をとる場合はもちろん比率も正の値しかとらない。さらに全体に対する一部分の比率のような場合（例えば総人口における男性人口の比率など）R は 0 以上 1 以下の値しかとらない。このような変数を線形のモデルで表現すると結果の解釈が非常に困難になる場合があり、あまり推奨できない。例えば

$$R_i = \beta X_i + u_i,\ i = 1,..,n$$

というような線形回帰モデルを考えたとき、最小二乗法による推定結果として $\hat{\beta} = 0.5$ というような値を得たとしよう。もし X_i の値のとりうる範囲になんの制約もない場合、X_i が 2 より大きい値をとった場合は R_i のあてはめ値が 1 より大きくなってしまうし、また X_i が 0 より小さい値をとってしまった場合は R_i のあてはめ値が 0 より小さくなってしまう。つまりあてはめ値が実際にはとりえない範囲になってしまう可能性がある。たとえ分析を行ったデータにおいてはこのような問題が発生しなかったとしても X_i の取りうる値の範囲で R_i のあてはめ値として実際にはあり得ない値になる可能性が潜在的にあるモデルというのは R_i の動きを表現するモデルとしてあまりよくないだろう。このような問題を解決するには、もし R_i のあてはめ値が 1 以上になったらすべて 1 とするなどがありえるがこれは非常に不自然であり、また推定も煩雑になり推奨できない。より自然な方法として説明変数がどのような値をとったとしても R_i はつねに 0 以上 1 以下の範囲に収まるようなモデルを考えるというものがある。例えば R_i が

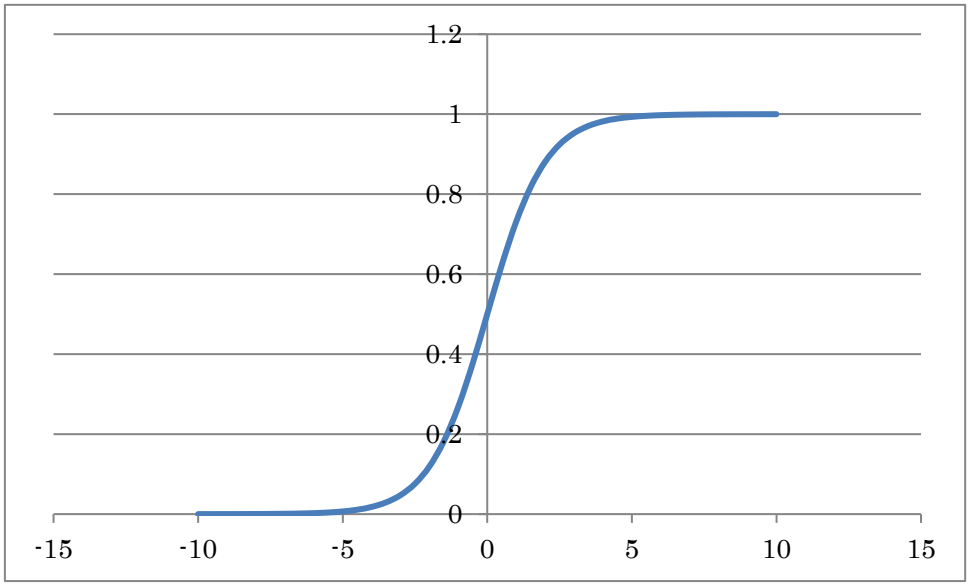
$$R_i = f(\beta X_i + u_i)$$

というように決定されるとしよう。ここで $f(x)$ は x のどのような値に対しても $0 < f(x) < 1$ であるような連続関数であるとする(実はさきほどの 1 以上なら 1、0 以下なら 0 というのもこのような関数で表す事ができるが、微分不可能な点が存在するので推定が煩雑になる)。このようにすれば上述した問題は起こらず、推定結果を自然に解釈できる。では $f(x)$ としてどのような関数形が考えられるであろうか？R_i が $0 < R_i < 1$ であるような場合 $f(x)$ の一つの候補として

$$f(x) = \frac{1}{1+\exp(-x)}$$

というものがある。この関数は x がどのような値をとっても常に 0 と 1 の間の数をとる。例えば x が∞に近づくほど $f(x)$ は 1 に近づくし、x が $-\infty$ に近づくほど $f(x)$ は 0 に近づく
この関数を図示したのが下の図である。ただしこの関数は、比率が 0.1 から 0.9 の間くら

いであれば直線でかなりよく近似できるのでもし比率の方のデータのほとんどが 0.1 から 0.9 の間くらいにあるようなデータであれば、比率を直接説明変数に回帰した場合と（係数が何倍かされるだけで係数の相対的な大きさの関係は）ほぼ同じになるであろう。



この関数を用いる利点の一つとして実はこのように R_i が決定される場合 R_i 自体はパラメーターに関して非線形であるが、$\log(R_i / (1 - R_i))$ がパラメーターの線形の関数となるので、それに対して最小二乗法によってパラメーターが推定できるという利点がある。これを見てみよう。今 i 番目の比率 R_i は

$$R_i = \frac{1}{1+\exp(-x_i)}, \qquad x_i = \alpha + \beta_1 X_{1i} + \ldots + \beta_K X_{Ki} + u_i$$

によって決定されるとしよう。ここで X_{ki}, k =1,.., K は説明変数、u_i は観測されない誤差項である。これらの式より

$$\begin{aligned}
[1+\exp(-x_i)]R_i = 1 &\Leftrightarrow R_i + R_i \exp(-x_i) = 1 \\
&\Leftrightarrow R_i \exp(-x_i) = 1 - R_i \\
&\Leftrightarrow \log(R_i) - x_i = \log(1 - R_i) \\
&\Leftrightarrow \log(R_i) - \log(1 - R_i) = x_i \\
&\Leftrightarrow \log\left(\frac{R_i}{1 - R_i}\right) = \alpha + \beta_1 X_{1i} + \ldots + \beta_K X_{Ki} + u_i
\end{aligned}$$

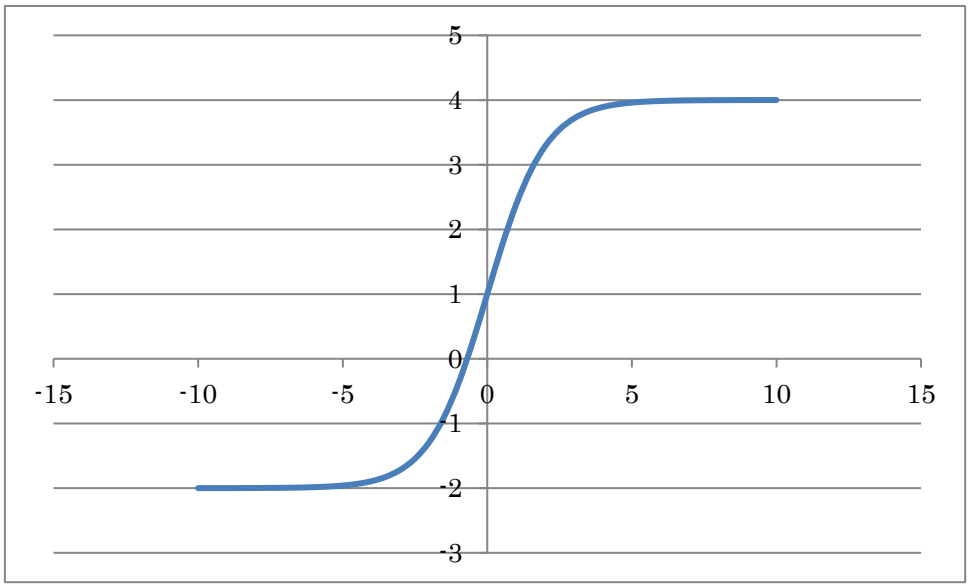
を得る。3 行目は両辺の対数を取ることにより得られる。これは、R_i そのものは線形モデルの被説明変数とするのはおかしいが、新しく $y_i = \log(R_i /(1 - R_i))$ という変数を計算して、これを被説明変数として用いれば、通常の線形回帰モデルになり最小二乗法によってパラメーターが推定できることを示している。この場合、例えば β_1 が正というのは X_{1i} が大きくなるほど R_i が 1 に近づくので X_{1i} は比率に対して正の影響を与えると解釈できるし、さらにどのような X_{1i} の範囲に対してもあてはめ値は 1 より小さくなるのでおかしなあてはめ値がで

る事もない。

より一般的にはある変数 S_i の取りうる値の範囲が $a < S_i < b$ のような場合(ここで a と b は既知であるとする)、S_i を線形のモデルで分析するのはやはり上記と同様の問題が生じるので具合が悪い。このような場合も上記のような変換が考えられる。具体的には例えば、

$$S_i = \frac{b-a}{1+\exp(-x_i)} + a\,, \qquad x_i = \alpha + \beta_1 X_{1i} + \ldots + \beta_K X_{Ki} + u_i$$

というように S_i が決定されるとすると、この S_i は $a < S_i < b$ の制約を自動的に満たすので上述の問題は起きない。この場合は x_i が大きくなると b に近づき、小さくなると a に近づく。下は $a = -2, b = 4$ としたときのこの関数の図である。



この時、

$$\begin{aligned}
&S_i[1+\exp(-x_i)] = (b-a) + a[1+\exp(-x_i)] \\
\Leftrightarrow\ &S_i + S_i\exp(-x_i) = b - a + a + a\exp(-x_i) \\
\Leftrightarrow\ &(S_i - a)\exp(-x_i) = b - S_i \\
\Leftrightarrow\ &\log(S_i - a) - x_i = \log(b - S_i) \\
\Leftrightarrow\ &\log\left(\frac{S_i - a}{b - S_i}\right) = x_i = \alpha + \beta_1 X_{1i} + \ldots + \beta_K X_{Ki} + u_i
\end{aligned}$$

であるので $y_i = \log((S_i - a)/(b - S_i))$ に対して線形モデルをあてはめ係数を最小二乗法で推定すればよいという事になる。