交互作用についてのメモ

2006.3.14最終修正

　もっとも単純な２つの２値的説明変数と１つの目的変数という場合（二元配置でもっとも単純な場合）を考える。
　２つの２値的説明変数を変数1と2とし、それぞれ0と１という値だけをとるとする。変数1が1であることは要因１が存在することを表し、0は要因１が存在しないことを表す。変数２についても同様である。目的変数は、生存率のような割合であることを想定し、0から1までの値しかとらないことにする。
　変数1=iで変数2=jのときの目的変数の期待値をpijとする。目的変数が生存率だとすると以下のような対応関係となる

要因１も要因２も存在しないときの生存確率をp00
要因１だけが存在するときの生存確率をp10
要因２だけが存在するときの生存確率をp01
要因１も要因２も存在するときの生存確率をp11

以下、目的変数が生存率であるとして進めるが、1と0の間の値という以外は生存率固有の性質は使っていないので、生存率でなくとも同様の議論が成り立つ量は多いと考えられる。

乗法的モデル

　乗法的モデルは、ある死亡要因だけが存在するときの生存率を（生存率1）、別の死亡要因だけが存在するときの生存率が（生存率2）であるとき、両要因が独立に作用したときの生存率は（生存率1）×（生存率2）であるとするものであり、多くの生存／死亡を扱う研究で仮定されている。乗法的モデルのもとでは、生存確率は以下のように表現できる。

要因１も要因２も存在しないときの生存確率をc
要因１だけが存在するときの生存確率をb×c
要因２だけが存在するときの生存確率をd×c
要因１も要因２も存在するときの生存確率をb×d×c

したがって、p00×p11=p10p01つまりp11=p10×p01／p00である。

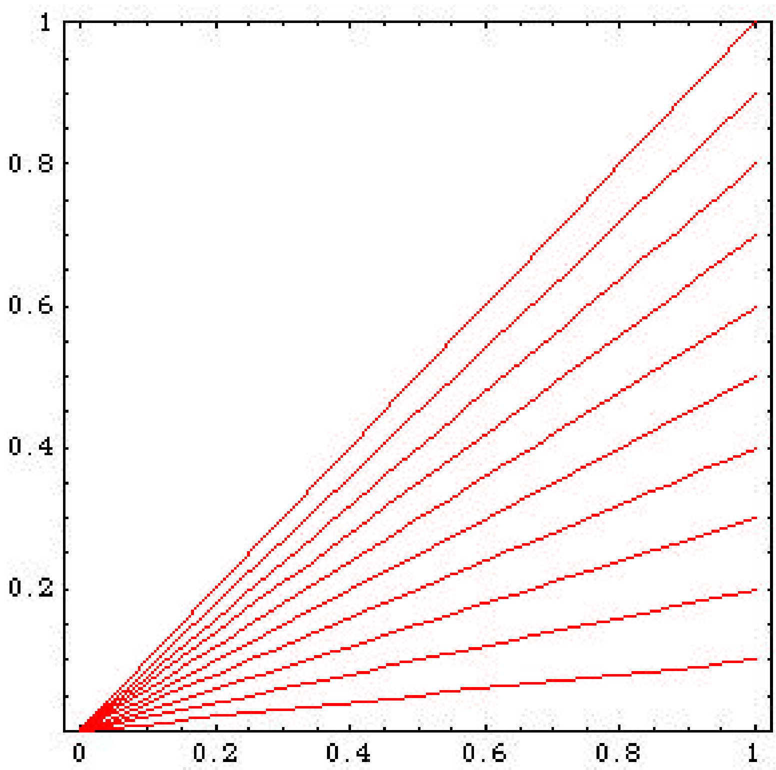
　乗法的モデルで死亡率に注目したら同じ関係が成り立つだろうか？

要因１も要因２も存在しないときの死亡確率は1-c
要因１だけが存在するときの死亡確率は1-b×c

要因２だけが存在するときの死亡確率は1-d×c
要因１も要因２も存在するときの死亡確率は1-b×d×c

このとき、(1-bc)(1-dc)-(1-c)(1-bdc)=c(1-b)(1-d)>0であるからさきほどの関係はもはや成り立たない。

　乗法的モデルでは、p11×p00=p10×p01であり、生存確率を行列（２×２）の形で書くと、”たすきがけ”した項の積が等しい。
　（ある条件のもとで要因２が存在しないときの生存確率、同じ条件で要因２が存在するときの生存確率）をx-y平面にプロットすることを考える。乗法的モデルのもとでは、要因2が存在すると生存確率はd倍になる。言い換えると、要因２の効果が一定の場合、要因２が存在するときの生存確率＝（d）・（他の条件は同じで要因２が存在しないときの生存確率）であるから、要因２の効果が等しい状態を線で結ぶと原点を通る直線となる（下図）。



別の表現をすれば、乗法的モデルでは、上図の同じ直線上にある点で表わされる状態は要因２の効果が等しいと考えているのである。

ロジスティックモデル

　生存率のような二項比率を目的変数としたとき、ロジスティック回帰などで使われているロジスティックモデルを適用することは自然に思える。

生存確率は以下のようになる。

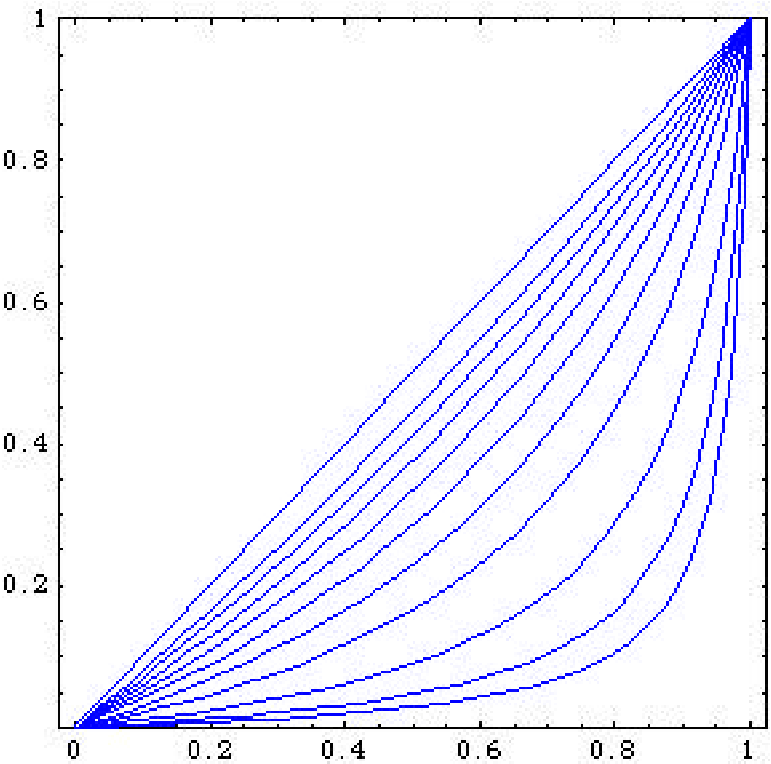
要因１も要因２も存在しないときの生存確率をC/(1+C)
要因１だけが存在するときの生存確率をBC/(1+BC)
要因２だけが存在するときの生存確率をDC/(1+DC)
要因１も要因２も存在するときの生存確率をBCD/(1+BCD)

　ロジットについては、以下のようになる。

要因１も要因２も存在しないときのロジットは、C/(1+C)を1/(1+C)で割ったものであるからＣ
要因１だけが存在するときのロジットはBC
要因２だけが存在するときのロジットはDC
要因１も要因２も存在するときのロジットはBCD

そこで、ロジット00×ロジット11＝ロジット01×ロジット10である。

　ロジスティックモデルでは、ロジット00×ロジット11＝ロジット01×ロジット10、つまり生存確率を行列（２×２）の形で書くと、生存確率そのものではなくロジットを”たすきがけ”した積が等しい。この関係を使って、要因２が存在するときの生存確率を縦軸に他の条件は同じで要因２が存在しないときの生存確率を横軸にとると、要因２の効果が等しい点は、原点を通る以下の図の曲線の上に並ぶ。



なお、ロジットを”たすきがけ”した積が等しいという関係はロジスティック回帰や対数線形モデルを考えると当然ではあるが、生存確率ではなく（1-生存確率）である死亡確率に注目しても成り立つ。また、p00、p10、p01が同じとき、ロジスティックモデルのp11の方が乗法的モデルのp11よりもいつでも小さいことになる。

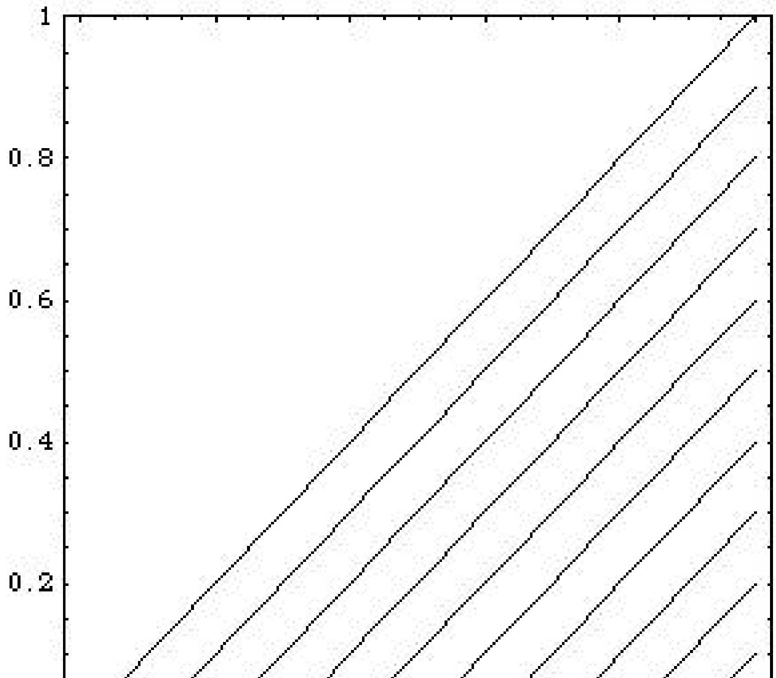
加法モデル

　分散分析や重回帰（線形の重回帰）で長く使用されて来たモデルである。生存率のような二項比率の目的変数に適用した場合、その値が0と1の間でなければいけないから、加法モデルの適用には制約があるが、想定によっては不適当ではないだろう。
　生存確率は以下のようになる。

要因１も要因２も存在しないときの生存確率をγ
要因１だけが存在するときの生存確率をγ+β
要因２だけが存在するときの生存確率をγ+δ
要因１も要因２も存在するときの生存確率をγ+β+δ

γは0と1の間である。βとδはいずれも負である。

　加法モデルでは、生存確率について、
（γ）＋（β＋δ＋γ）＝（β＋γ）＋（δ＋γ）
つまり、p00+p11=p10+p01という関係が成り立っている。すなわち、加法モデルでは、行列（2×2）の形で書くと、生存確率を”たすきがけ”にみて足した和が等しい。また、ここから、要因２が存在するときの生存確率＝（定数）＋（他の条件は同じで要因２が存在しないときの生存確率）となっているときに、要因２の効果が等しいとみなしていることになる。要因２が存在するときの生存確率を縦軸に他の条件は同じで要因２が存在しないときの生存確率を横軸にとると、要因２の効果が等しい点は、下図のように45度の角度で右上がりの直線の上に並ぶ。



また、p00、p10、p01が同じなら、加法モデルのp11の方が乗法的モデル

のp11よりもいつでも小さいことになる。

対数線形モデル

　ロジスティックモデルと同じであるから省略する。

比例ハザードモデル

　Coxの比例ハザードモデルと呼ばれることもある。比例ハザードモデルは、生存時間などの分析に使われることが普通であるから、ここで考えているような一定時間経過後の生存と死亡を見ていると考えられるようなケースに使うことはあまりないかもしれない。だが、比例ハザードモデルが想定している関係をここで考えている状況に適用することは単純である。

　比例ハザードモデルでは、個体のハザードは
（基準ハザード）・exp（独立変数の一次式）
という形で表現される。（独立変数の一次式）を**βｚ**とすると、生存率との関係は、基準ハザードをh0として、

$$S(t) = e^{-h_0 \, e^{\beta z}}$$

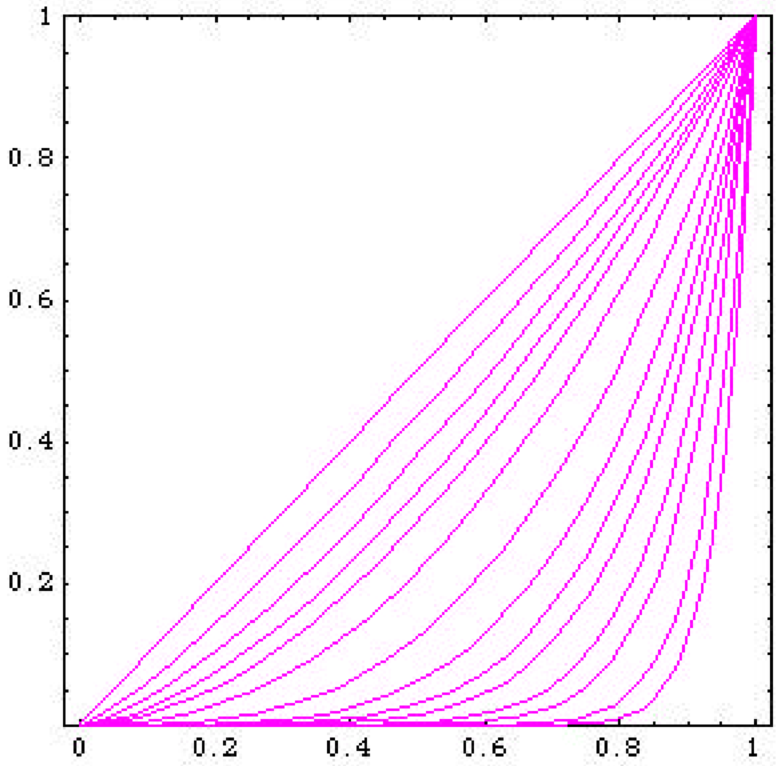
である。いま、**βｚ**が具体的にはα＋β1・x1＋β2・x2であるとする。

（0､0）では、$S(t) = e^{-h_0 \, e^{\alpha}}$
（0､1）では、$S(t) = e^{-h_0 \, e^{\alpha + \beta 1}}$
（1､0）では、$S(t) = e^{-h_0 \, e^{\alpha + \beta 2}}$
（1､1）では、$S(t) = e^{-h_0 \, e^{\alpha + \beta 1 + \beta 2}}$

それぞれの対数は、

（0､0）では、$\log S(t) = -h_0 \;\; e^{\alpha}$
（0､1）では、$\log S(t) = -h_0 \;\; e^{\alpha + \beta 1}$
（1､0）では、$\log S(t) = -h_0 \;\; e^{\alpha + \beta 2}$
（1､1）では、$\log S(t) = -h_0 \;\; e^{\alpha + \beta 1 + \beta 2}$

となる。つまり、乗法モデルで”たすきがけ”の積が、加法モデルで”たすきがけ”の和が等しかったのに対して、比例ハザードモデルでは”たすきがけ”の対数の積が等しいのである。以下の図はこれまでのモデルと同様に要因２の効果が等しい点を連ねたものである。

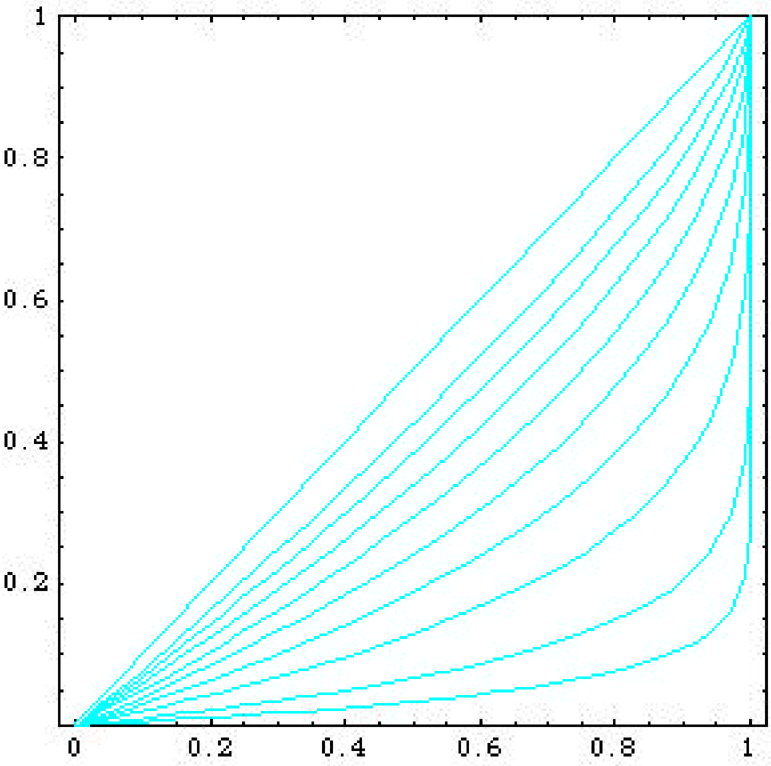


complimentary log-logモデル

　complimentary log-logモデルでは、期待値（確率）と独立変数の一次式の関係は

（独立変数の一次式）$= \log\{-\log(1-\pi)\}$

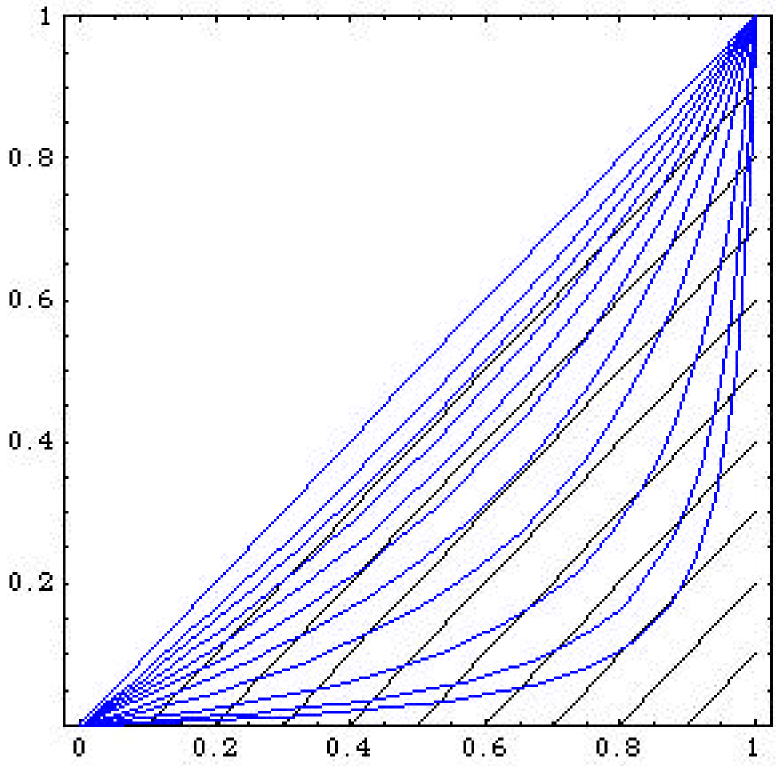
であり、”たすきがけ”にみたとき、比例ハザードモデルでは生存率の対数の積に対して成り立っていた関係が、complimentary log-logモデルでは（1-生存率）の対数の積に対して成り立つことがわかる。



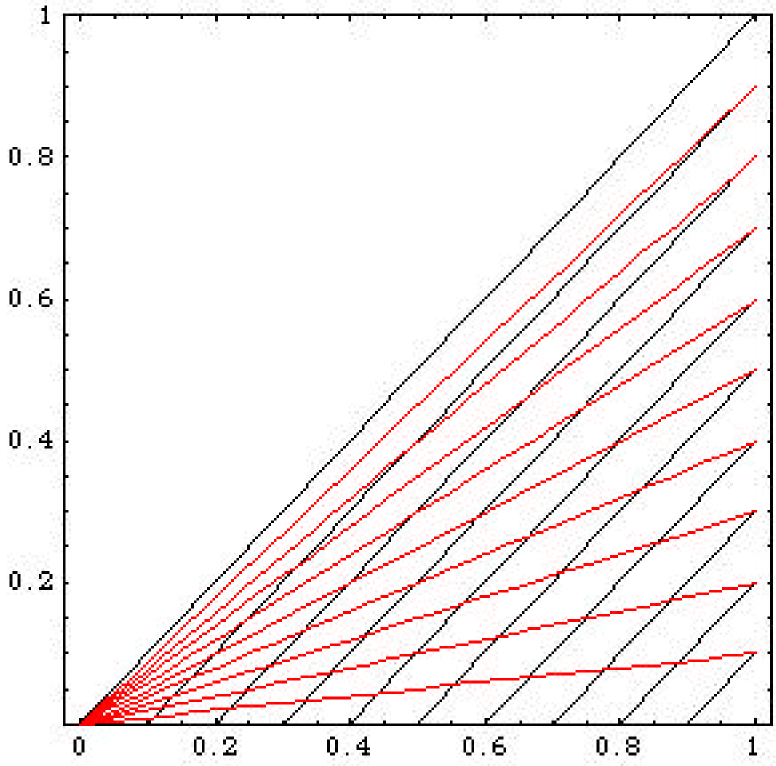
以下の図はこれまでのモデルと同様に要因２の効果が等しい点を連ねたものである。

モデルの比較

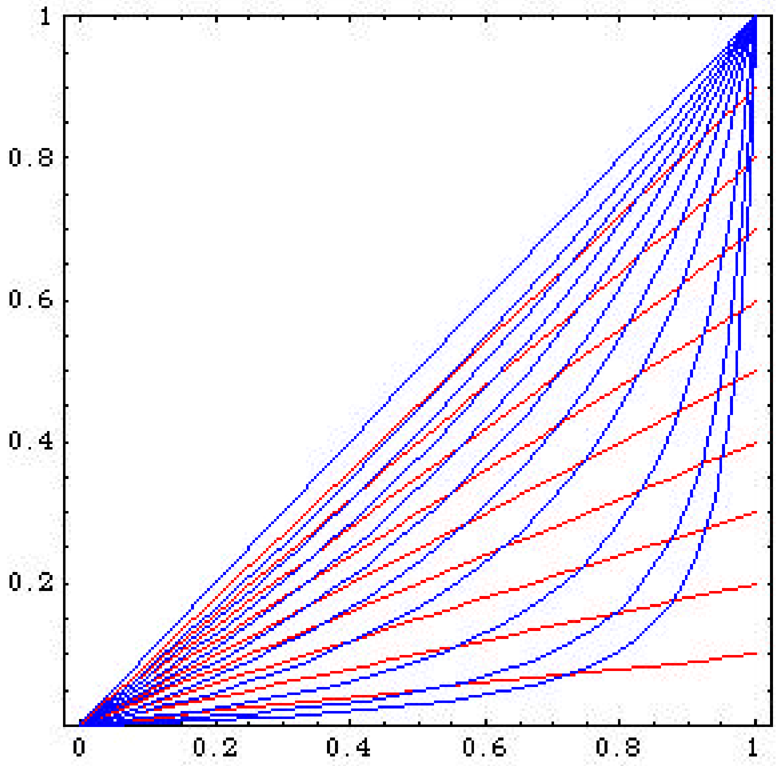
　ここでは、要因２の等効果線とでも言うべきものをプロットした図を見てきた。複数のモデルについて作ったプロットを重ねたとき、片方のモデルでの等効果線が、もう片方のモデルでの複数の等効果線と交わっているなら、片方のモデルでは効果が等しいと見なされる要因の効果がもう片方のモデルでは等しくなく変化しているみなされることを意味する。どちらのモデルを適用するかにより、要因の効果の大きさの推定が異なってくるわけである。
　下図は加法モデルとロジスティックモデルのプロットを重ねたものである。もともとの形が大きく違うので当然だが、１本の線が他のモデルの多くの線と交差している。



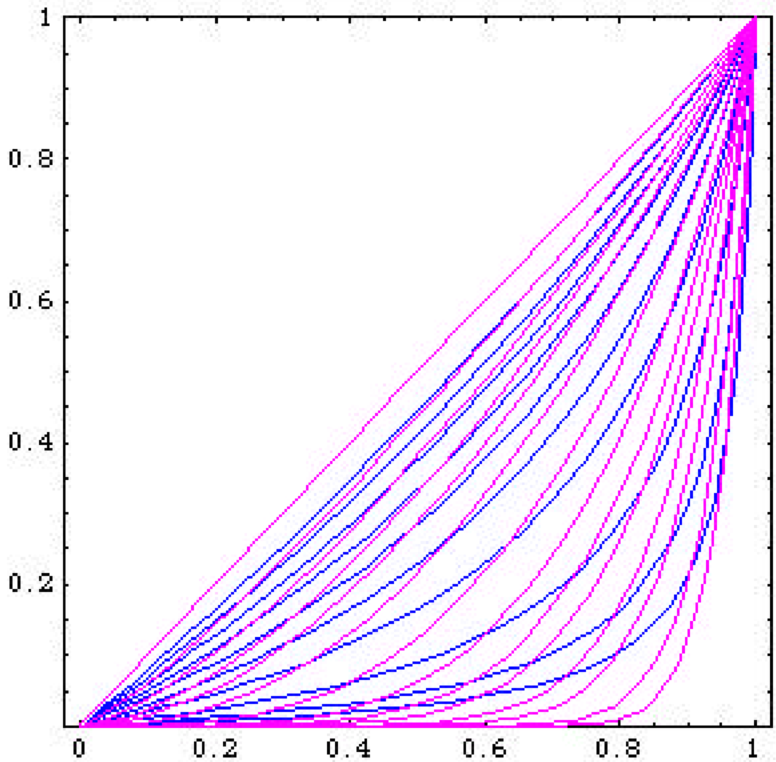
下図のように加法モデルは乗法モデルとも大きく異なることがわかる。



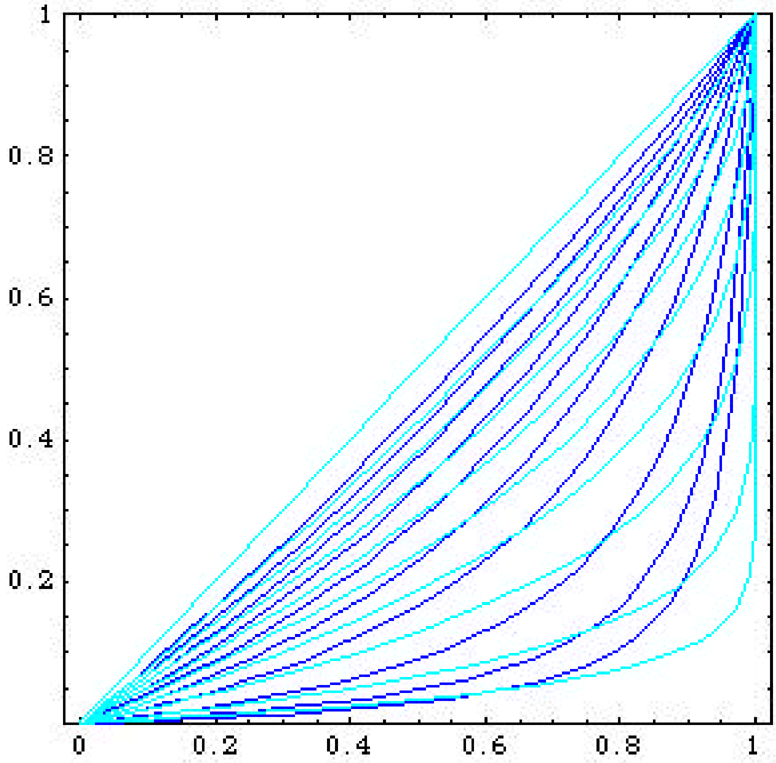
それぞれのプロットから予想されることだが、乗法モデルとロジスティックモデルのちがいも大きいことが下図からわかる。



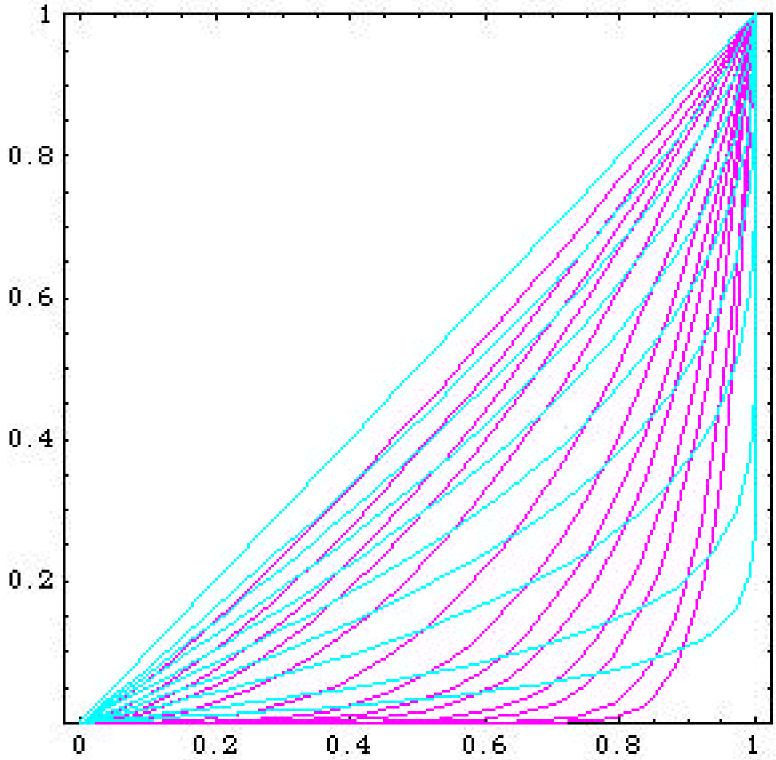
単独のプロットでは、ロジスティックモデルと比例ハザードモデルは似ているようにも見えるが、要因の効果が大きいときにはかなり大きなちがいがあることがわかる（下図）。



下図に示すロジスティックモデルとcomplimentary log-logモデルの場合も、ロジスティックモデルvs比例ハザードモデルの比較とよく似ている。要因の効果が大きいときにはかなり大きなちがいがあることがわかる。



　一応、比例ハザードモデルとcomplimentary log-logモデルも比べておく。要因の効果が弱いときを除くとかなり大きなちがいがあることがわかる。



意味

どのモデルを適用するかにより、要因の効果の大きさの推定値が異なるわけであるが、ある意味では当然の結果でもある。複数の要因が関与していると考えるとき、要因の効果の大きさ（回帰係数や分散分析のような分析法なら主効果）は、要因の効果の間に特定の関係（足し算とか掛け算とかロジットとか）を仮定せずには推定できないので、データだけから自然に導かれるものではない。

そこで、”なぜ他のモデルではなくあるモデルを使うのか”がいつも必要であることになる。研究の歴史の中で、特定のモデルを使うことに合理的で明確な理由がある研究分野やテーマではとくに問題は起こらないだろう。だが、そうではないときには、『他の要因の効果をコントロールしたときの、その要因の効果』とか『他の要因の効果を除いたときの、その要因の効果』について述べようとすると、いつでも分析に使われた特定のモデルの使用を合理化する理由が必要である（統計的方法の教科書では、モデルの決定論的部分はそれぞれの研究分野により自ずから明らかと考えられている傾向があると思う）。たとえば、発芽率のデータの分析に、ロジスティックモデルを使うか、加法モデルを使うか、生存率みたいなものだから乗法モデルなのか、（それとも他のモデルか）によって、検出される効果やその意味はちがってくる。

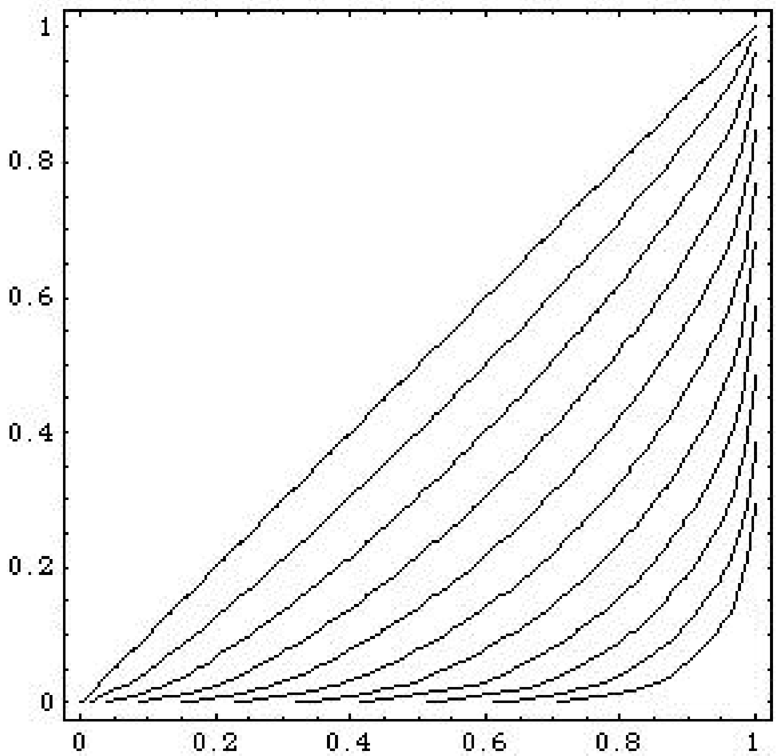
さて、そのようなことを考えなくていい例外的な場合がある。与えられたデータともっともよく合っているモデルを見つければいいときである。このときには、あっているモデルを選べばいいので、あっているだけで十分である。しかし、実はそのようなケースはあまり多くはないだろう。

他の要因の効果をコントロールしたり除いたりしたいなら、あっているモデルを探すだけではすまない。コントロールしたり除いたりするには、要因の効果の間の関係を決めておく必要があるからである。また、もっともよく合うモデルを選ぶと、同様なデータではちがうモデルを選ぶことがあるだろう。そのとき、回帰係数や主効果の大きさのようなパラメーターは比べることができない（もっともよく合うモデルを探せばいいのなら、比較できなくてもいっこうにかまわないことになる）。

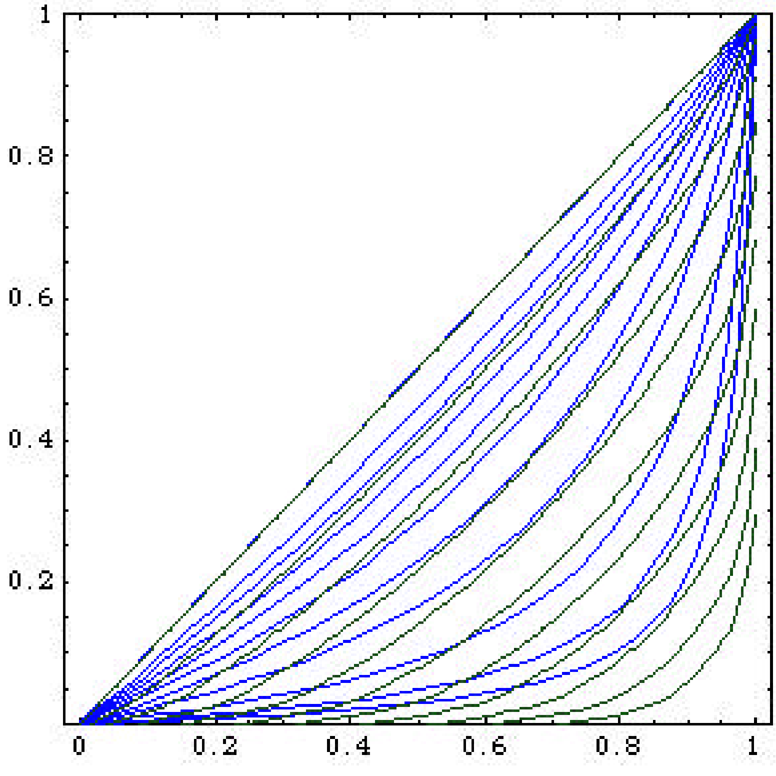
ここで扱ったモデルは一端である。たとえば、べき乗変換のファミリーを考えないのは不自然かもしれない。これを考えると、目的変数のλ乗＝説明変数の一次式、という決定論的部分を持つモデルが使える。このときには、λのとりうる値により無限個のモデルを考えることになる。

おまけ．変換ファミリーのうち角度変換

角度変換（逆正弦平方根変換）の場合は以下のようになる。



これは一見ロジスティックモデルによく似ている。だが、重ねてみると



”曲がり方”がことなり、交差が多数生じていることがわかる。