2025/04/22

藪友良

カウントデータ

カウントデータ(count data)とは、ある事象が一定期間に生じた回数を記録したデータになります。回数は非負の整数(0、1、2、…)だけをとります(概念上、マイナスの回数はありません)。たとえば、1 か月間に病院に行った回数、1 日の交通事故件数、ある女性の子供の数などが該当します。カウントデータは、一般に、小さな値を取ることが多く、また、0 の値を取ることが多いデータです。仮に、大きな値を取り、0 の値を取らないなら、通常の連続変数として扱うことができます。本稿では、被説明変数がカウントデータとした代表的モデルとして、ポアソン回帰モデルと負の二項分布モデルを紹介します[1][2]。

本稿は、藪友良『入門 実践する計量経済学』(2023 年、東洋経済新報社)の補足資料になります。

線形モデル

被説明変数Yがカウントデータであっても、通常の線形モデルとして推定することは可能です。

$$Y = \alpha + \beta X + u$$

このとき、限界効果(X が 1 単位変化したときのYの変化量)はβとなり、その解釈も容易になります。線形モデルは有用である一方、その欠点として 2 点が挙げられます。

第 1 に、カウントデータYは 0 を含むことが多く、その場合、被説明変数として、対数$\ln(Y)$を用いることができないという点です。なお、この点はカウントデータの対数をとっていないなら問題はありません。

第 2 に、線形モデルは予測値に負の値が生じるという欠点です。これは予測値$\hat{Y}$がXの線形関数、つまり、

$$\hat{Y} = \hat{\alpha} + \hat{\beta} X$$

であり、説明変数Xが大きかったり、小さかったりすると、予測値が負の値になるからです。カウントデータは非負の整数ですから、予測値が負となるというのはおかしな性質です。

これらの問題を避けるためには、カウントデータであることを明示的に扱ったモデルを考える必要があります。これがポアソン回帰モデルと負の二項分布モデルになります。以下では、ポアソン分布を紹介した後、期待値の対数が線形モデルとしたポアソン回帰モデル、また、ポアソン回帰モデルを拡張した負の二項分布モデルを紹介します。

---

[1] 分析方法の詳細に関心がある読者は、以下の書籍を参照してください。Cameron, A. Colin, and Pravin K. Trivedi. "Regression Analysis of Count Data," Cambridge University Press, Cambridge, U.K.

[2] 近年では、貿易における重力モデルを推定するために、ポアソン回帰モデルが使われるようになっています。貿易額は大きな値をとる変数であり、ポアソン回帰とは無関係に思われます。この点に関心がある読者は、サポートウェブサイトの追加資料「貿易における重力モデル」を参照してください。

ポアソン分布

ポアソン分布は、稀にしか発生しない事象の発生回数を表す離散確率分布となります。その例としては、ある航空会社の 1 年間の事故件数、ある病気の 1 日の新規感染者数などが挙げられます。

ポアソン分布(Poisson distribution)

確率変数Yが値yをとる確率は次のように表せます。

$$P\{Y = y\} = \frac{e^{-\mu}\mu^{y}}{y!}$$

値yは非負の整数(0、1、2、…)であり、$\mu > 0$は**強度**と呼ばれるパラメータです。記号!は階乗を表し、たとえば、$0! = 1$、$1! = 1$、$2! = 2 \times 1$、$3! = 3 \times 2 \times 1$です。

このとき、ポアソン確率変数Yの期待値は強度μとなり、分散もまた強度μとなります。

$$E[Y] = \mu$$

$$V(Y) = \mu$$

つまり、強度μは稀な事象の平均回数であり、回数の分散にもなっています(期待値と分散の導出は補足を参照してください)。

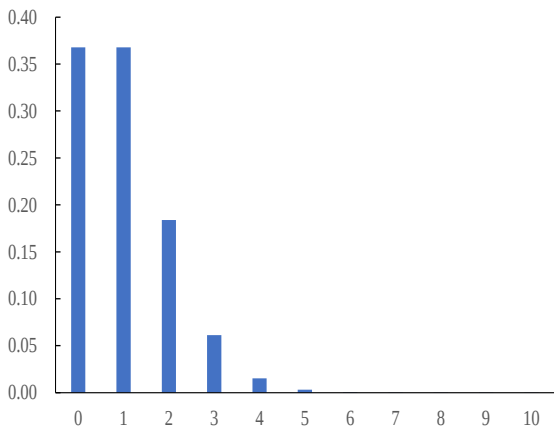
図 1(a)では、強度$\mu = 1$としたポアソン分布を描いています。図をみると、$y = 0$、もしくは$y = 1$の確率が高く、それ以降は確率が小さくなります。これを数式から確認しましょう。$\mu = 1$のとき$1^{y} = 1$となることに注意すると、確率は次のようになります。

$$P\{Y = y\} = \frac{e^{-1}}{y!}$$

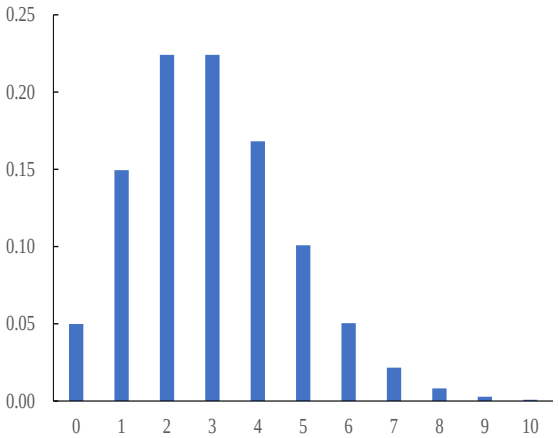
ここで$0! = 1$と $1! = 1$から、$y = 0$と$y = 1$ の確率はともに$e^{-1} = 0.368$です。また、$y = 2$なら確率$e^{-1}/2! = 0.184$、$y = 3$なら確率$e^{-1}/3! = 0.061$となります。

図 1：ポアソン分布

(a) $\mu = 1$



(b) $\mu = 3$



これに対して、図 1(b)では、$\mu = 3$としたポアソン分布を描いています。図をみると、yは 2 と 3 の確率が高くなります。これらの図から、期待値と分散が強度μに依存して変わることが理解できます。

例 1：新規感染者

ある病気の 1 日の新規感染者数の平均が 1 人だったとしましょう。このとき、強度を$\mu = 1$と設定できますから、新規感染者数の分布は図 1(a)のようになります。感染者数が 0 人となる確率は 37%、1 人となる確率は 37%、2 人となる確率は 18%、3 人となる確率は 6%程度です。平均が 1 人であっても、0 人となったり、2 人以上となったりすることが頻繁に生じます。

<u>ポアソン回帰モデル</u>

ポアソン分布では、強度μは定数と仮定されますが、強度μは説明変数Xに依存しているかもしれません。たとえば、学生の欠席回数は、年齢や性別などによっても変わるかもしれません。強度が説明変数に依存する可能性を考慮したのがポアソン回帰モデルになります。以下では簡単化のため、説明変数が確率変数ではないとして議論します。

ポアソン回帰モデル(Poisson regression model)

確率変数Yはポアソン分布とし、強度μは次のように設定されます。

$$\mu = e^{\alpha+\beta X}$$

ポアソン分布では、強度μはYの期待値となりますから、

$$E[Y] = e^{\alpha+\beta X}$$

となります。指数関数は 0 より大きな値しかとらないので、期待値$E[Y]$は 0 より大きな値になります。また、分散も強度であり、次のようになります。

$$V(Y) = e^{\alpha+\beta X}$$

ここで、分散はXの値に依存して変わるため、これは**不均一分散**を意味しています。

なお、期待値の式の対数をとると、

$$\ln(E[Y]) = \alpha + \beta X$$

となります。つまり、ポアソン回帰モデルは、期待値の対数を被説明変数とした対数線形モデルとしても解釈ができます。なお、Yは 0 の値をとることもありますが、Yの期待値は 0 より大きいため、期待値の対数を定義することが可能です。

<u>ポアソン回帰モデルの係数の解釈</u>

モデルは$\ln(E[Y]) = \alpha + \beta X$ であるため、係数βは対数線形モデルの解釈と同じです(6.3.2 節参照)。この点を確認しましょう。Xが$X + 1$に変化したとき、$E[Y]$は$E[Y']$に変化するとし

ます。上記のモデルから、次の式が成立します。

$$\ln(E[Y']) - \ln(E[Y]) = (\alpha + \beta(X+1)) - (\alpha + \beta X) = \beta$$

これをβについて解くと、

$$\beta = \ln(E[Y']) - \ln(E[Y]) = \ln\left(1 + \frac{E[Y'] - E[Y]}{E[Y]}\right) \approx \frac{E[Y'] - E[Y]}{E[Y]}$$

となります(巻末付録 A 参照)。記号≈は近似で等しいことを表します。これはXが 1 単位変化すると、期待値$E[Y]$が 100×β%変化することを意味します。

上式は、Yの変化が小さいときの近似ですが、Yの変化が大きいとき近似は不正確となります。このとき、近似を使わないでも、$\beta = \ln(E[Y']) - \ln(E[Y])$が成立するため、

$$e^{\beta} = \frac{E[Y']}{E[Y]}$$

となります(この式が正しいことは両辺の対数をとったら明らかでしょう)。つまり、Xが 1 単位変化すると、期待値が$100 \times (e^{\beta} - 1)$%変化することを意味します。$Y$の変化が 30%くらいまでは近似でも問題ありませんが、それを超えたら、こちらを使ってみることをお勧めします(補足参照)。

線形モデルとの比較に関心があるなら、通常の限界効果(Xが 1 単位変化したときのYの変化量)を推定することも可能です。$E[Y] = e^{\alpha+\beta X}$において、$Z = \alpha + \beta X$と置くと、合成関数の微分の公式によって、限界効果は次の式になります。

$$\frac{dE[Y]}{dX} = \frac{de^{Z}}{dZ}\frac{dZ}{dX} = e^{\alpha+\beta X}\beta$$

限界効果は、Xの値によって値が変わります。実証分析では、プロビットやロジットと同様に、平均限界効果が掲載されます。ここで、予測値を$\hat{Y}_i = e^{\hat{\alpha}+\hat{\beta}X_i}$と定義すると、**平均限界効果**(average marginal effect)は次のように計算できます。

$$\frac{1}{n}\sum_{i=1}^{n}\frac{dE[Y_i]}{dX_i} = \hat{\beta}\frac{1}{n}\sum_{i=1}^{n}e^{\hat{\alpha}+\hat{\beta}X_i} = \hat{\beta}\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i$$

ここで、Y_iと$\hat{Y}_i$の平均は同じため、平均限界効果は$\hat{\beta}\,\bar{Y}$としても計算できます。

最尤推定

ポアソン回帰モデルは非線形モデルであり、通常の最小 2 乗法によって推定できません。しかし、ここでYはポアソン分布であることが分かっているため、**最尤法**によってパラメータを推定することができます(最尤法は 12.3 節参照、ポアソン回帰の尤度は補足参照)。

ポアソン分布は、期待値と分散がともに強度μとしているため、ポアソン回帰モデルでも、期待値と分散は同じと仮定されます。

$$V(Y) = E[Y] = e^{\alpha+\beta X}$$

これは非現実的仮定です。実際、カウントデータの多くは、$V(Y) > E[Y]$という傾向がみられます。この現象は**過剰分散**(overdispersion)と呼ばれます。

ポアソン分布の仮定が誤っていたとしても、$E[Y] = e^{\alpha+\beta X}$ が成立していれば、最尤推定量はパラメータの一致推定量となることが知られています。ただし、ポアソン分布の仮定が誤っていると、通常の標準誤差は、本当の標準誤差を過小評価する傾向があるため、**ロバスト標準誤差**を用いることが推奨されます。

<u>負の二項分布モデル</u>

ポアソン分布の拡張として、負の二項分布モデルがあります。負の二項分布モデルでは、追加的なパラメータθが導入されることで、期待値と分散が異なる可能性が許容されます。

負の二項分布モデル(negative binominal model)

ポアソン分布の強度μは、次のように設定されます。

$$\mu = Ve^{\alpha+\beta X}$$

ここで、Vは正の値を取る確率変数であり、その期待値は 1、分散は$1/\theta$と仮定されます(ただし、$\theta > 0$とする)[3]。パラメータθが∞であれば、Vは常に値 1 をとるため、負の二項分布はポアソン分布と同じになります。

このとき、Yの期待値と分散は、それぞれ

$$E[Y] = e^{\alpha+\beta X}$$

$$V(Y) = e^{\alpha+\beta X}\left(1 + \frac{e^{\alpha+\beta X}}{\theta}\right)$$

となり、期待値と分散が異なる可能性が許容されています。なお、$e^{\alpha+\beta X}$とθは正の値をとりますから、$e^{\alpha+\beta X} > 0$です。

これらの式から、2 つの点が明らかです。第 1 に、負の二項分布では、過剰分散($V(Y) > E[Y]$)が制約として課されている点です。このため、過剰分散ではないなら、負の二項分布は適切ではないかもしれません。第 2 に、θが∞なら、$V(Y) = E[Y]$となる点です。このため、負の二項分布はポアソン回帰を内包したモデルといえます。

カウントデータは通常の線形モデルでも推定できますが、結果の頑健性を調べるためにも、ポアソン回帰モデルと負の二項分布モデルによる推定も合わせて行うことが推奨されます[4][5]。

---

[3] 厳密には、確率変数Vはガンマ分布 Gamma(θ、θ)に従うと仮定されます。

[4] カウントデータに 0 が多く含まれている場合、ゼロ過剰(zero inflated)ポアソン回帰とゼロ過剰負の二項分布モデルを用いることができます。たとえば、ゼロ過剰ポアソン回帰では、確率pでカウントが 0 となり、確率$1-p$でカウントがポアソン回帰から決定されるとします。

[5] Stata では、ポアソン回帰は poisson コマンド、負の 2 項分布は nbreg コマンドで実行できます。最後に r を入れるとロバスト標準誤差、irr を入れると係数 β の代わりに、e^{β}が計算されます。また、限界効果に関心があるなら、margins, dydx(説明変数)とすれば計算できます。Stata では、θ ではなく、alpha= 1/θ が推定されます。R のプログラムに関しては、松浦寿幸氏のウェブサイトに説明がまとめられています。https://note.com/toshi_matsuura/n/n75177561bd30

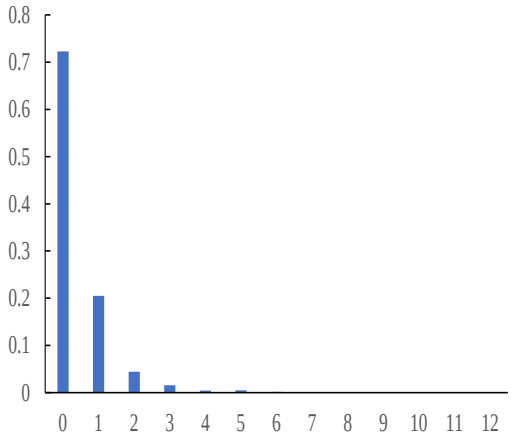
例 2：逮捕回数の決定要因は何か

Wooldridge『Introductory Econometrics』では、カリフォルニア生まれの男性 2725 人分のデータ(CRIME1)を用いて、逮捕回数の決定要因が分析されています。これらの男性は、1960 年もしくは 1961 年に生まれており、1986 年より前に逮捕歴があります。

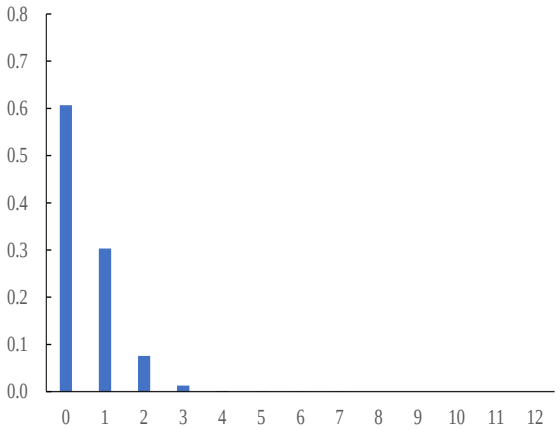
図 2(a)では、1986 年の逮捕回数の相対頻度を描いており、逮捕回数は 0 回から 12 回の間で分布しています。逮捕回数 0 回が約 72%、1 回が約 20%となっています。図 2(b)では、強度$\mu = 0.5$としたポアソン分布を描いており、ポアソン分布が逮捕回数の分布を近似できていることがわかります。

図 2: 逮捕回数とポアソン分布

(a) 逮捕回数の相対頻度



(b) 強度$\mu = 0.5$としたポアソン分布



被説明変数は 1986 年の逮捕回数であり、説明変数は、有罪率(1986 年より前の逮捕のうち有罪となった割合)、投獄期間(1986 年に何カ月投獄されていたか)、所得(1986 年の所得)、黒人ダミーとします。表 1 では、線形モデル、ポアソン回帰、負の二項分布モデルからの推定結果を示しています。どの推定でも係数の符号や有意性は同じです。過去の有罪率が高いと、罰則の厳しさを恐れるからか、1986 年の逮捕回数が減少します。また、投獄期間が長いと、その間は悪いことができないため、逮捕回数が減少しています。また、所得が上がると逮捕回数は減少し、黒人であると(白人に比べて)逮捕回数が増加します。

線形モデルと他のモデルでは、係数の意味が異なることに注意が必要です。たとえば、線形回帰モデルでは、投獄期間が 1 か月増えると 0.029 だけ逮捕回数が減少しています。これに対し、ポアソン回帰では、投獄期間が 1 か月増えると 8.1%だけ逮捕回数が減少します。線形モデルとの比較のため、ポアソン回帰の係数-0.081から限界効果をもとめてみます。逮捕回数の平均が 0.4044 なので、限界効果は$-0.033(= -0.081 \times 0.4044)$です。これは線形モデルの係数とほぼ同じ値です。

同様に、線形モデルでは、黒人は白人より 0.288 回だけ逮捕回数が増えるとしています。これに対し、ポアソン回帰の係数 0.5114 は、近似を使うと、黒人は 51.4%だけ逮捕回数が増えることを意味します。しかし、Yの変化は大きいため、近似は不正確であると疑われま

す。実際、正しい計算は$0.672(=e^{0.514}-1)$であり、0.514 よりずいぶん大きくなります。これは黒人であると白人より 67.2%だけ逮捕回数が多いことがわかります。また、ポアソン回帰の限界効果は 0.2078(=0.514×0.4044)であり、線形モデルの係数に近い値になっています。

表 1: 推定結果

	線形モデル	ポアソン回帰	負の二項分布
過去の有罪率	-0.136 ***	-0.389 ***	-0.450 ***
	(0.034)	(0.100)	(0.102)
投獄期間	-0.029 ***	-0.081 ***	-0.088 ***
	(0.005)	(0.020)	(0.022)
所得	-0.002 ***	-0.009 ***	-0.009 ***
	(0.000)	(0.001)	(0.001)
黒人ダミー	0.288 ***	0.514 ***	0.510 ***
	(0.058)	(0.093)	(0.092)
定数項	0.547 ***	-0.501 ***	-0.483 ***
	(0.028)	(0.062)	(0.064)
1/θ			0.989
			(0.139)
AIC	6746.98	4556.17	4362.36
OBS	2745	2745	2745

注）カッコ内はロバスト標準誤差を用いた。また、***は 1%有意、**は 5%有意、*は 10%有意を表す。OBS はサンプルサイズとなる。

表 1 をみると、ポアソン回帰と負の二項分布の係数はほぼ同じ値になっています。これはポアソン分布の推定量は一致性を持っていることから妥当な結果でしょう。また、負の二項分布において、1/θは 0.989 ですから、θは逆数 1.011 となります。θは小さな値ですから、過剰分散が生じており、負の二項分布が妥当なモデルとわかります。実際、AIC をみると、負の二項分布が最も小さな値となり、当てはまりが良いモデルとわかります。6.4.3 節では、AIC が小さいほど良いモデルであったことを思い出してください。

補足

ポアソン分布は確率の和が 1 となる

ここで確率の和が 1 となることを確認します。まず、

$$e^{\mu} = \sum_{y=0}^{\infty} \frac{\mu^y}{y!}$$

となります[6]。この結果から、

$$\sum_{y=0}^{\infty} \frac{e^{-\mu}\mu^y}{y!} = e^{-\mu} \sum_{y=0}^{\infty} \frac{\mu^y}{y!} = e^{-\mu}e^{\mu} = 1$$

ポアソン分布の期待値と分散

Yがポアソン分布に従うとし、期待値と分散がそれぞれμとなることを証明します。まず、期待値は次のようになります。

$$\begin{aligned} E[Y] &= \sum_{y=0}^{\infty} y \frac{e^{-\mu}\mu^y}{y!} \\ &= \sum_{y=1}^{\infty} y \frac{e^{-\mu}\mu^y}{y!} \\ &= \mu e^{-\mu} \sum_{y=1}^{\infty} \frac{\mu^{y-1}}{(y-1)!} \end{aligned}$$

ここで$k = y - 1$と定義すると、上式は

$$\mu e^{-\mu} \sum_{k=0}^{\infty} \frac{\mu^k}{k!}$$

となり、ここで$e^{\mu} = \sum_{k=0}^{\infty} \frac{\mu^k}{k!}$から期待値は$\mu$となります。

次に 2 乗の期待値は次のように展開できます。

$$\begin{aligned} E[Y^2] &= \sum_{y=0}^{\infty} y^2 \frac{e^{-\mu}\mu^y}{y!} \\ &= \sum_{y=0}^{\infty} (y(y-1) + y) \frac{e^{-\mu}\mu^y}{y!} \end{aligned}$$

6　$e^{\mu x}$をxに関してマクローリン展開すると、次のようになります。

$$e^{\mu x} = 1 + \frac{1}{1!}\mu x + \frac{1}{2!}\mu^2 x^2 + \frac{1}{3!}\mu^3 x^3 + \cdots$$

式展開では、$\frac{de^{\mu x}}{dx} = \mu e^{\mu x}$、$\frac{d^2 e^{\mu x}}{dx^2} = \mu^2 e^{\mu x}$、$\frac{d^3 e^{\mu x}}{dx^3} = \mu^3 e^{\mu x}$を用いました。上式を$x = 1$で評価すると次式が得られます。

$$e^{\mu} = 1 + \frac{1}{1!}\mu + \frac{1}{2!}\mu^2 + \frac{1}{3!}\mu^3 + \cdots = \sum_{y=0}^{\infty} \frac{\mu^y}{y!}$$

$$= \sum_{y=0}^{\infty} y(y-1)\frac{e^{-\mu}\mu^{y}}{y!} + \sum_{y=0}^{\infty} y\frac{e^{-\mu}\mu^{y}}{y!}$$

右辺第 2 項は期待値μであり、右辺第 1 項は

$$\sum_{y=0}^{\infty} y(y-1)\frac{e^{-\mu}\mu^{y}}{y!} = \sum_{y=2}^{\infty} y(y-1)\frac{e^{-\mu}\mu^{y}}{y!}$$

$$= \mu^2 e^{-\mu}\sum_{y=2}^{\infty}\frac{\mu^{y-2}}{(y-2)!}$$

となります。ここで、$k = y - 2$と定義すると、上式はμ^2となります。以上から、2 乗の期待値は$E[Y^2] = \mu^2 + \mu$です。

分散は 2 乗の期待値から期待値の 2 乗を引いたものですから、次式が得られます。

$$V(Y) = E[Y^2] - E[Y]^2$$

$$= (\mu^2 + \mu) - \mu^2 = \mu$$

<u>ポアソン回帰の対数尤度</u>

ポアソン回帰モデルでは、

$$P\{Y_i = y_i\} = \frac{e^{-\mu}\mu^{y_i}}{y_i!}$$

としています(ただし、$\mu = e^{\alpha+\beta X_i}$)。上式の対数をとると、次のようになります。

$$y_i(\alpha + \beta X_i) - \left(e^{\alpha+\beta X_i}\right) - \ln(y_i!)$$

ここで$\ln(y!)$は、パラメータ(α、β)に依存していないため無視することができます。よって、最尤法は対数尤度である次式を最大にするようにパラメータを決めます。

$$\sum_{i=1}^{n}\{y_i(\alpha + \beta X_i) - e^{\alpha+\beta X_i}\}$$

対数尤度をαに関して偏微分して 0 と置くと

$$\frac{\partial}{\partial\alpha}\sum_{i=1}^{n}\{y_i(\alpha + \beta X_i) - e^{\alpha+\beta X_i}\} = \sum_{i=1}^{n}\{y_i - e^{\alpha+\beta X_i}\} = 0$$

となり、さらにβに関して偏微分して 0 と置くと次の式が得られます。

$$\frac{\partial}{\partial\beta}\sum_{i=1}^{n}\{y_i(\alpha + \beta X_i) - e^{\alpha+\beta X_i}\} = \sum_{i=1}^{n}\{y_i - e^{\alpha+\beta X_i}\}X_i = 0$$

残差を$y_i - e^{\alpha+\beta X_i}$と考えると、最初の式は残差の和が 0、2 番目の式は残差と説明変数の積和が 0 を意味しています。なお、両式を満たすパラメータ(α、β)は最尤推定量($\hat{\alpha}_{ML}$、$\hat{\beta}_{ML}$)となりますが、解析的には解けないため、数値探索法で見つけることになります。

<u>対数近似の正確</u>

係数の解釈を行うとき、次の近似を行いました。

$$\ln\left(1+\frac{E[Y']-E[Y]}{E[Y]}\right)\approx\frac{E[Y']-E[Y]}{E[Y]}$$

この近似は変化率が何%までであれば正確といえるでしょうか。ここで変化率をεとして、$\ln(1+\varepsilon)$とεを図示しました。図 3 をみると、εが 0.2 や 0.3 くらいまでは正確ですが、それを超えると差が大きくなっていくことがわかります。このため、変化率が 0.2 や 0.3 を超えていたら、近似を使わない計算を調べてみることが大事だと思います。

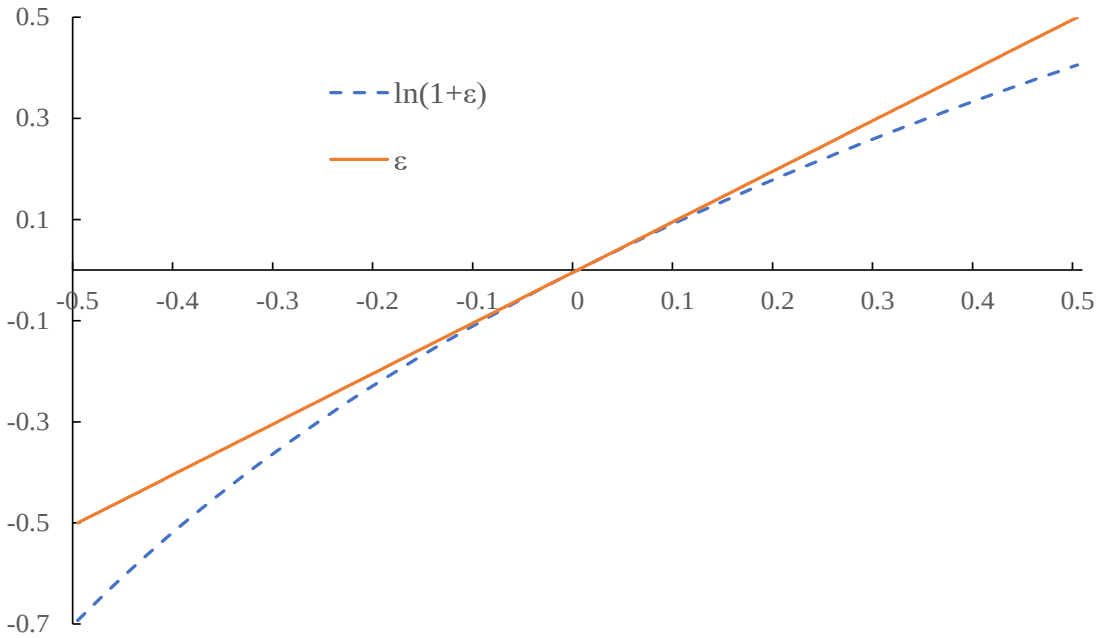



図 3: 対数差の近似は正確か？