カウントデータモデルI

モデルの紹介
係数の解釈
推定

1

カウントデータ

- カウントデータは「何回起きたか？」を記録するもの。
- よって**非負の整数**（0,1,2,3,….）のみを値として取ります。
- 例としては、、、
  - 一年間に行った旅行の回数（個人レベル）
  - 一日に起こる自殺の件数（県レベル）
  - 国際的な紛争の件数（国レベル、一年）
  - サッカーの得点（チームレベル、一試合）
  - 飲食店の新規出店数（市レベル、一年）

2

- カウントデータのモデル化には、線形回帰モデルがしばしば使われます。
- 確かに、ケースによっては、カウントデータに線形回帰モデルを当てはめることに問題は無いかもしれません。
- しかし、ケースによっては、線形回帰モデルの最小二乗推定量は、一致性や有効性を持たないことがあります。
- 従って、カウントデータを分析する場合は、カウントデータのために特別にデザインされたモデルを使う方が安全といえます。

3

カウントデータに線形モデルを当てはめることの問題点

Y_i：従属変数（カウントデータ）$X_{ij}(j=1,\dots k)$：説明変数

- 最も単純なモデリングは線形モデル：$Y_i=\beta_0+\beta_1X_{i1}+\cdots+\beta_kX_{ik}+U_i$

で、最小二乗法で推定です。

- このモデリングの弱点は、線形確率モデルの弱点と似ています。具体的には、、、、
  - Y_iは非負の値しか取らない。よって、$E(Y_i|X_{i1},\dots,X_{ik})$は、$X_{ij}(j=1,\dots k)$の値の組み合わせがいかなるものであっても、正の値を取るべき。
  - しかし、$\hat{Y}_i^{OLS}=\hat{\beta}_0^{OLS}+\hat{\beta}_1^{OLS}X_{i1}+\cdots+\hat{\beta}_k^{OLS}X_{ik}$は、何らかの$X_{ij}(j=1,\dots k)$の値の組み合わせにおいて負の値を取る。
  - 負の予測値、、、

4

- Y_iが正の値しか取らないとき、Y_iに対数を取って

$$lnY_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + U_i$$

とモデル化することが良くあります。

- しかし、Y_iがカウントデータの場合、このアプローチは適切なものになり難い、、、
- なぜなら、カウントデータでは、$Y_i = 0$という観測値が数多くあるのが普通だからです。
  - 対数は$Y > 0$のときだけ定義されます。
  - そのため、このアプローチを使うと、$Y = 0$の観測値は分析から抜け落ちます（missing value扱いです）。
  - $Y = 0$の観測値がシステマティックに分析から抜け落ちてしまう、、、これでは$\boldsymbol{Y}$と$\boldsymbol{X}$の関係を正しく分析できないかもしれません。

5

- 対数変換によって$Y = 0$の観測値が分析から抜け落ちてしまうのを防ぐために、Yに1を足して

$$\ln(Y_i + 1) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + U_i$$

とモデル化することがよくあります。

- しかし、係数の解釈に問題が生じるかもしれません。
  - なぜなら上のモデルの係数をあたかも下のモデルの係数

$$\ln(Y_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + U_i$$

のように解釈していいのは、$Y = 0$の観測値があまりないときだけです（Wooldridge, J.M.（2012）*Introductory Econometrics*, 5th edition, South-Western の193ページ参照）。

  - カウントデータでは、$Y = 0$という観測値が数多くあるのが普通です。

6

ポワソン分布（Poisson distribution）

- カウントデータを回帰モデル化する際、「ポワソン分布」が重要な役割を果たします。
- まずポワソン分布の定義を与えます。

離散型確率変数Yの確率質量関数が

$$Pr(Y = h) = exp(-\lambda)\frac{\lambda^h}{h!}, \qquad h = 0,1,2, \ldots$$

で与えられるとき、Yが従う分布をパラメータλのポワソン分布という。

- パラメータλは、この分布に従うポワソン確率変数の期待値かつ分散になります。すなわち

$$E(Y) = VAR(Y) = \lambda \quad \text{（期待値と分散は同じ！）}$$

7

- 前のページの!は、階乗（かいじょう）[英：factorial]を表しています。

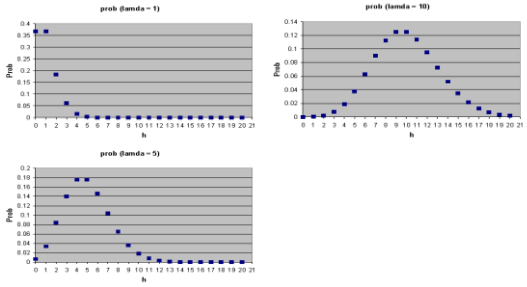
$$n! = n \times (n-1) \times (n-2) \times \cdots \times 1$$

$0! = 1$（これは定義です）
$1! = 1$
$2! = 2 \times 1$
$3! = 3 \times 2 \times 1$

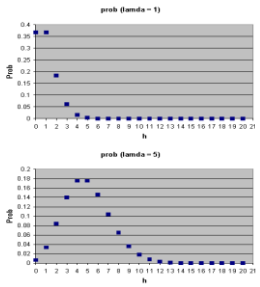
8

ポワソン分布：$Pr(Y = h) = exp(-\lambda)\frac{\lambda^h}{h!}$

- パラメーターであるλに値を入れると、分布の形が決まります。



9



（1）λ（期待値）が大きくなると、確率が集中しているところが右にシフト。

（2）λ（期待値）が大きくなると、散らばりも大きくなる（平均=分散）

（3）λ（期待値）が大きくなると、ゼロの確率が急激に小さくなる。

（4）λ（期待値）が大きくなると、正規分布に近づく。

10

ポワソン分布の特性と現実のデータ

- ポワソン分布では、期待値＝分散。これを等分散（equidispersion）仮定と言います。
- しかし、実際のデータでは、標本分散の方が標本平均より大きいことが多い（過分散: overdispersion）。
  - 実際のデータで、標本分散の方が標本平均より小さいこと（過少分散：underdispersion）はレア。
- ポアソン分布では、λ（平均）が大きくなると、ゼロの確率が急激に小さくなる。
- しかし、現実のデータでは、ポワソン分布が予測するゼロよりも、より多くのゼロが観測されることが多い。

11

ポワソン回帰モデル（Poisson regression model）

- このモデルを考える際には、これまでのように誤差項が、、、とかはありません。
- ポワソン分布の期待値を説明変数でモデル化して終わりです。
- まずポワソン分布を考えます。

$$Pr(Y = h) = exp(-\lambda)\frac{\lambda^h}{h!}, \qquad h = 0,1,2,\dots$$

- 添字iをつけましょう。期待値のパラメーターλにもです。

$$Pr(Y_i = h) = exp(-\lambda_i)\frac{\lambda_i^h}{h!}, \qquad h = 0,1,2,\dots$$

  - 個人個人で持っているポワソン確率変数の期待値が異なる、って感じですね。

12

- そして、Y_iの期待値であるでλ_iが、説明変数$X_{ij}(j=1,\dots,k)$の関数になっている、と考えます。
- λ_iは説明変数の値によって変わる、 $\lambda_i(X_{i1},\dots,X_{ik})$ってことですね。
- 言い換えれば、説明変数$X_{ij}(j=1,\dots,k)$を条件とした、Y_iの条件付き期待値$E(Y_i|X_{i1},\dots,X_{ik})$を考えるということです。
- 線形回帰モデルがやってることは$E(Y_i|X_{i1},\dots,X_{ik})$のモデル化です。
- なので、ポワソン回帰モデルも線形回帰モデルもやってることは同じですね。

13

- 線形回帰モデルでは、

$$E(Y_i|X_{i1},\dots,X_{ik})=\beta_0+\beta_1X_{i1}+\cdots+\beta_kX_{ik}$$

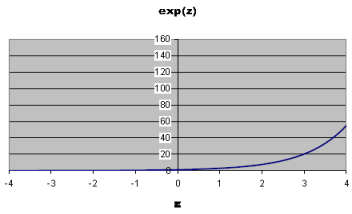
- しかし、このスペシフィケーションは、カウントデータの分析ではちょっと都合が悪いです。
- なぜなら、説明変数$X_{ij}(j=1,\dots,k)$の値の何らかの組み合わせにおいて、$E(Y_i|X_{i1},\dots,X_{ik})$の値が負になってしまう、、、
- Y_iは非負の値しか取らないのに、$E(Y_i|X_{i1},\dots,X_{ik})$が負の値を取りうる、っていうのはモデルとして問題です。
- よってポワソン回帰モデルでは、こうします：

$$E(Y_i|X_{i1},\dots,X_{ik})=exp(\beta_0+\beta_1X_{i1}+\cdots+\beta_kX_{ik})$$

14

$$E(Y_i|X_{i1},\dots,X_{ik})=exp(\beta_0+\beta_1X_{i1}+\cdots+\beta_kX_{ik}) \quad (式1)$$

- こうしておけば、説明変数$X_{ij}(j=1,\dots,k)$の値がいかなるものであっても、$E(Y_i|X_{i1},\dots,X_{ik})$は正の値を取ります。



- 従って、 Y_iの予測値$\hat{Y}_i$が、常に正になることが保証されます。

15

$$E(Y_i|X_{i1},\dots,X_{ik})=exp(\beta_0+\beta_1X_{i1}+\cdots+\beta_kX_{ik}) \quad (式1)$$

- 実は、このスペシフィケーションは、皆さんが知っているモデルに似ています。
- （式1）の両辺に対数をとれば、、、

$$lnE(Y_i|X_{i1},\dots,X_{ik})=\beta_0+\beta_1X_{i1}+\cdots+\beta_kX_{ik}$$

なので、皆さんがご存知の対数線形モデル

$$E(\ln Y_i|X_{i1},\dots,X_{ik})=\beta_0+\beta_1X_{i1}+\cdots+\beta_kX_{ik}$$

と期待値オペレーターの位置を除けば一緒です。

16

ポワソン回帰モデル：まとめ

$$Pr(Y_i = h|X_{i1}, \dots, X_{ik}) = exp\left(-\lambda_i(X_{i1}, \dots, X_{ik})\right)\frac{\lambda_i^h}{h!}, \qquad h = 0,1,2, \dots$$

$$\lambda_i(X_{i1}, \dots, X_{ik}) = E(Y_i|X_{i1}, \dots, X_{ik}) = exp(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik})$$

以上がポワソン回帰モデルの根幹。

$$Pr(Y_i = 0|X_{i1}, \dots, X_{ik}) = exp\left(-exp(\boldsymbol{X_i\beta})\right)$$

$$Pr(Y_i = 1|X_{i1}, \dots, X_{ik}) = exp\left(-exp(\boldsymbol{X_i\beta})\right) \times exp(\boldsymbol{X_i\beta})$$

$$Pr(Y_i = 2|X_{i1}, \dots, X_{ik}) = exp\left(-exp(\boldsymbol{X_i\beta})\right) \times \frac{[exp(\boldsymbol{X_i\beta})]^2}{2}$$

……

- 確率質量関数は

$$f_{Y_i}(y_i|X_{i1}, \dots, X_{ik}) = exp\left(-exp(\boldsymbol{X_i\beta})\right)\frac{[exp(\boldsymbol{X_i\beta})]^{y_i}}{y_i!}$$

- 結合確率質量関数は

$$f_{Y_1,\dots,Y_N}(y_1, \dots, y_N|\boldsymbol{X_1}, \dots, \boldsymbol{X_k}) = \prod_{i=1}^{N} exp\left(-exp(\boldsymbol{X_i\beta})\right)\frac{[exp(\boldsymbol{X_i\beta})]^{y_i}}{y_i!}$$

- 尤度関数は

$$L(\boldsymbol{\beta}|Y_1 = y_1, \dots, Y_N = y_N, \boldsymbol{X_1}, \dots, \boldsymbol{X_k}) = \prod_{i=1}^{N} exp\left(-exp(\boldsymbol{X_i\beta})\right)\frac{[exp(\boldsymbol{X_i\beta})]^{y_i}}{y_i!}$$

- 対数尤度関数は

$$lnL(\boldsymbol{\beta}|Y_1 = y_1, \dots, Y_N = y_N, \boldsymbol{X_1}, \dots, \boldsymbol{X_k}) = \sum_{i=1}^{N}[y_i \cdot (\boldsymbol{X_i\beta}) - exp(\boldsymbol{X_i\beta}) - lny_i!]$$

条件付き期待値の推計値・条件付き確率分布の推計

- $\boldsymbol{\beta}$の推定値$\widehat{\boldsymbol{\beta}}$が得られたら、

$$\hat{Y}_i = \hat{E}(Y_i|X_{i1}, \dots, X_{ik}) = exp(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_k X_{ik})$$

のようにY_iの条件付き期待値の推計値を計算できます。

- 同様に、Y_iの条件付き確率分布も推計できます。

$$\widehat{Pr}(Y_i = 0|X_{i1}, \dots, X_{ik}) = exp\left(-exp(\boldsymbol{X_i\widehat{\beta}})\right)$$

$$\widehat{Pr}(Y_i = 1|X_{i1}, \dots, X_{ik}) = exp\left(-exp(\boldsymbol{X_i\widehat{\beta}})\right) \times exp(\boldsymbol{X_i\widehat{\beta}})$$

……

等分散の仮定は問題になる？

- ポワソン分布では、等分散（期待値＝分散）の仮定が置かれていることを説明しました。
- しかし、実際のデータでは、標本分散の方が標本平均より大きいことが多い（過分散: overdispersion）。
- 仮定が満たされていない → ポワソン回帰モデルは役に立たない？
- 一般的に言って、分布の仮定が正しくないとき、最尤推定量は一致性をもちません。
  - 例えば、二項プロビット・ロジット、順序ロジット・プロビット、インターバル回帰モデル、多項ロジット、条件付きロジット、これから学習予定のトービットモデル、すべてそうです。

- しかし、ポワソン回帰モデルは例外です。

たとえポワソン分布の仮定が正しくなかったとしても、
ポワソン分布の仮定に基づいた$\boldsymbol{\beta}$の最尤推定量は一致性を持つ。

参照：Gouriéroux, C., Monfort, A., and Trongnon, A. （1984）"Pseudo Maximum Likelihood Methods: Applications to Poisson Models," *Econometrica*, 52, 701-720.

- すなわち、

$\boldsymbol{\beta}$のポワソン最尤推定量は、
分布のミススペシフィケーションに対して
頑健性（robustness）を持つ。

ということです。

ノート：統計手法の仮定している条件が満たされていないとします。そのような場合でもほぼ妥当な結果を与えるとき、その手法は「頑健（robust）である」とか「頑健性を持つ」といいます。

21

- ポワソン回帰モデルの$\boldsymbol{\beta}$を最尤法で推定、ただし「ポワソン分布の仮定が正しい」とは必ずしも仮定しない

このアプローチを

ポワソン疑似最尤推定
（Poisson Quasi Maximum Likelihood Estimation）

と言います。

- データは過分散していてポワソン分布に従っていないかもしれない、、、そんなときでも、ポワソン回帰モデルを最尤推定して問題ない、と言うことになります。

- 従って、ポワソン回帰モデルは見た目以上にパワフルなモデルです。

22

- $\boldsymbol{\beta}$のポワソン疑似最尤推定量が一致性を持つための条件は、、、

Y_iの条件付き期待値$E(Y_i|X_{i1},\dots,X_{ik})$が正しくモデル化されていることのみ。

- ポワソン疑似最尤推定について、詳しくは、

Wooldridge, J.M.（2010）*Econometric Analysis of Cross Section and Panel Data* (2nd edition), The MIT press, Cambridge, Massachusetts.

のp.727~728 を参照してください。

23

ポワソン疑似最尤推定を使うときの注意

（1）必ずロバスト標準誤差を使うこと

考えられる二つのケース。

ケース1：実際にポワソン分布だった。

最尤推定量は一致性を持ちます。標準誤差の計算も、スタンダードな標準誤差の計算の仕方で問題ありません。

ケース2：ポワソン分布ではなかった。

最尤推定量は一致性を持ちます（ポワソン疑似最尤推定量）。しかし、スタンダードな標準誤差の計算の仕方は、正しい標準誤差を与えないことが知られています。

妥当な標準誤差の計算方法はいくつかありますが、「**ロバスト標準誤差**」を使うのが最も簡単です。

ノート：分析の際は、常にケース2であることを想定するのが安全です。

24

（2）推計確率を信用しすぎない

- $\boldsymbol{\beta}$のポワソン疑似最尤推定量は一致性を持ちます。
- 従って、 $X_{i1}, \dots, X_{ik}$が与えられたときの Y_iの条件付き期待値の推計値は信頼に値するものと言えます。
- ただし、分布の仮定が誤っているかもしれないので、$X_{i1}, \dots, X_{ik}$が与えられたときの Y_iの条件付き確率分布の推計はあまり信用しない方が良いかもしれません。

25

（3）従属変数は非負の整数でなくても構わない。

- この点は注意点というより、知っておくといいことです。
- ポワソン回帰モデルは「非負の整数」をモデル化したいときのみにしか使えない、と考えがちです。
- しかし、

> たとえポワソン分布の仮定が正しくなかったとしても、
> ポワソン分布の仮定に基づいた$\boldsymbol{\beta}$の最尤推定量は一致性を持つ。

- これより

> ポワソン回帰モデルは、従属変数が非負の値しか取らないのであれば、
> 連続型変数のモデリングにも使える

ことが知られています。

26

- 特に、

「従属変数が、連続的＋非負の値しか取らない＋ゼロの値をよくとる」

ときにポワソン疑似最尤推定はとても有効です。

- 有名な応用例は、

Santos Silva, J.M.C. and Tenreyro, Silvana (2006), The Log of Gravity, *The Review of Economics and Statistics*, 88(4), pp. 641-658.

この論文は、ポワソン疑似最尤推定アプローチを使って、二国間の貿易のフローを説明しています。

- 多くの二国間で貿易のフローはゼロです。

- 詳しくは、http://personal.lse.ac.uk/tenreyro/LGW.html を参照してください。

27

ポワソン回帰モデルの結果の解釈

$$E(Y|X_1, \dots, X_k) = exp(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k)$$

- 説明変数$X_j (j = 1, \dots, k)$が連続型とします。このとき、

$$\frac{\partial E(Y|X_1, \dots, X_k)}{\partial X_j} = exp(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k)\beta_j$$

> 他の条件を一定として、X_jが1単位増えたとき、
> Yが平均的にはどれだけ変化するか

- さらに、

$$\beta_j = \frac{\left(\frac{\partial E(Y|X_1, \dots, X_k)}{\partial X_j}\right)}{E(Y|X_1, \dots, X_k)} = \frac{\partial ln(E(Y|X_1, \dots, X_k))}{\partial X_j}$$

> 他の条件を一定として、X_jが1単位増えたとき、
> Yは平均的には$100\beta_j$%変化する

> もしX_jが対数をとったものであれば、β_jは弾力性

> 従属変数が$\ln(Y)$のときの係数の解釈の仕方と同じ

28

- もし説明変数がダミー変数なら、、、、

他の条件を一定として、X_jが**1**単位増えたときのYの平均的な変化は

$$E(Y|X_1,\dots X_j=1,\dots,X_k)-E(Y|X_1,\dots X_j=0,\dots,X_k)$$
$$=exp(\beta_0+\beta_1X_1+\cdots+\beta_j+\cdots+\beta_kX_k)-exp(\beta_0+\beta_1X_1+\cdots+0+\cdots+\beta_kX_k)$$

- X_jが**1**単位増えたときの Yの平均的な％変化は

$$100\times\frac{E(Y|X_1,\dots X_j=1,\dots,X_k)-E(Y|X_1,\dots X_j=0,\dots,X_k)}{E(Y|X_1,\dots X_j=0,\dots,X_k)}=100\times\left(exp(\beta_j)-1\right)$$

29

ポワソン回帰モデル：まとめ

$$Pr(Y_i=h|X_{i1},\dots,X_{ik})=exp(-\lambda_i(X_{i1},\dots,X_{ik}))\frac{\lambda_i^h}{h!},\qquad h=0,1,2,\dots$$

$$\lambda_i(X_{i1},\dots,X_{ik})=E(Y_i|X_{i1},\dots,X_{ik})=exp(\beta_0+\beta_1X_{i1}+\cdots+\beta_kX_{ik})$$

- たとえポワソン分布の仮定が正しくなかったとしても、ポワソン分布の仮定に基づいた$\boldsymbol{\beta}$の最尤推定量は一致性を持つ（ポワソン疑似最尤推定）。
- ただし、ポワソン分布の仮定が正しい場合は、スタンダードな標準誤差の計算の仕方は正しい標準誤差を与えない。
- よって必ずロバスト標準誤差を使うのがよい。
- 係数の読み方は、対数線形モデルのときとほぼ同じ。

30