Multinomiales Logit-Modell

Ulrich Kohler, Uni-Mannheim

Outdated version. Not longer supported

Inhaltsverzeichnis

1 MNLM und BLM

Für eine abhängigen Variable mit 3 Kategorien lassen sich drei binäre Logit-Modelle berechnen. Ein MNLM berechnet diese drei Regressionsmodelle gleichzeitig.

▷ Beispiel:

Variable Parteiidentifikation mit den Kategorien SPD, CDU und Keine Parteiidentifikation.

$$\ln\left[\frac{\Pr(\text{SPD}|\mathbf{x})}{\Pr(\text{CDU}|\mathbf{x})}\right] = \mathbf{x}\mathbf{b}_{\text{SPD vs. CDU}} \tag{1}$$

$$\ln\left[\frac{\Pr(\text{CDU}|\mathbf{x})}{\Pr(\text{Keine}|\mathbf{x})}\right] = \mathbf{x}\mathbf{b}_{\text{CDU vs. Keine}} \tag{2}$$

$$\ln\left[\frac{\Pr(\text{SPD}|\mathbf{x})}{\Pr(\text{Keine}|\mathbf{x})}\right] = \mathbf{x}\mathbf{b}_{\text{SPD vs. Keine}} \tag{3}$$

Dabei gilt:

$$\ln\left[\frac{\Pr(\text{SPD}|\mathbf{x})}{\Pr(\text{CDU}|\mathbf{x})}\right] + \ln\left[\frac{\Pr(\text{CDU}|\mathbf{x})}{\Pr(\text{Keine}|\mathbf{x})}\right] = \ln\left[\frac{\Pr(\text{SPD}|\mathbf{x})}{\Pr(\text{Keine}|\mathbf{x})}\right] \tag{4}$$

▷ Beispiel:

```
. tab pi

         pi |      Freq.     Percent        Cum.
------------+-----------------------------------
        SPD |        586       20.89       20.89
        CDU |        521       18.57       39.47
      Keine |      1,698       60.53      100.00
------------+-----------------------------------
      Total |      2,805      100.00
```

Aus (4) folgt:

$$\ln\left[\frac{\Pr(\text{SPD}|\mathbf{x})}{\Pr(\text{CDU}|\mathbf{x})}\right] + \ln\left[\frac{\Pr(\text{CDU}|\mathbf{x})}{\Pr(\text{Keine}|\mathbf{x})}\right] = \ln\left[\frac{\Pr(\text{SPD}|\mathbf{x})}{\Pr(\text{Keine}|\mathbf{x})}\right] \tag{5}$$

$$= \mathbf{x}\mathbf{b}_{\text{SPD vs. CDU}} + \mathbf{x}\mathbf{b}_{\text{CDU vs. Keine}} = \mathbf{x}\mathbf{b}_{\text{SPD vs. Keine}}$$

Sind also die Koeffizienten von zwei der drei möglichen Vergleiche bekannt, können die Koeffizienten des dritten Vergleichs aus den bekannten Koeffizienten abgeleitet werden.

In der Praxix kann MNLR nicht mit mehreren BLM berechnet werden, da jedes BLM auf einer anderen "Stichprobe" beruht.

```
. tab1 M*

-> tabulation of Mspdcdu

    Mspdcdu |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        521       47.06       47.06
          1 |        586       52.94      100.00
------------+-----------------------------------
      Total |      1,107      100.00

-> tabulation of Mcduno

     Mcduno |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,698       76.52       76.52
          1 |        521       23.48      100.00
------------+-----------------------------------
      Total |      2,219      100.00
```

```
-> tabulation of Mspdno
```

Mspdno	Freq.	Percent	Cum.
0	1,698	74.34	74.34
1	586	25.66	100.00
Total	2,284	100.00	

2 Wahrscheinlichkeitsmodell

Gegeben die AV y mit J nominalen Kategorien und $\Pr(y = m|\mathbf{x})$ die Wahrscheinlichkeit, bei gegebenen $\mathbf{x}$ die Kategorie m zu beobachten.

1. Es wird angenommen $\Pr(y = m|\mathbf{x})$ ist eine Funktion von $\mathbf{x}\mathbf{b}_m$, wobei $\mathbf{b}_m$ für jedes m unterschiedlich ist.
2. Um sicher zu stellen, dass $\Pr(y = m|\mathbf{x})$ nicht negativ werden kann verwenden wir $\exp(\mathbf{x}\mathbf{b}_m)$.
3. Um sicher zu stellen, dass $\sum_{m=1}^{J} \Pr(y = m|\mathbf{x}) = 1$ teilen wir $\exp(\mathbf{x}\mathbf{b}_m)$ durch $\sum_{j=1}^{J} \exp(\mathbf{x}\mathbf{b}_j)$.

Somit ist

$$\Pr(y = m|\mathbf{x}) = \frac{\exp(\mathbf{x}\mathbf{b}_m)}{\sum_{j=1}^{J} \exp(\mathbf{x}\mathbf{b}_j)} \tag{6}$$

3 Identifikation

Gleichung (6) ist nicht identifiziert, da mehrere $\mathbf{b}_m$ dieselben Wahrscheinlichkeiten erzeugen:

$$\begin{aligned}
\Pr(y = m|\mathbf{x}) &= \frac{\exp(\mathbf{x}\mathbf{b}_m)}{\sum_{j=1}^{J}\exp(\mathbf{x}\mathbf{b}_j)} \times \frac{\exp(\mathbf{x}\tau)}{\exp(\mathbf{x}\tau)} \\
&= \frac{\exp(\mathbf{x}\mathbf{b}_m + \mathbf{x}\tau)}{\sum_{j=1}^{J}\exp(\mathbf{x}\mathbf{b}_j) + \mathbf{x}\tau} \\
&= \frac{\exp(\mathbf{x}(\mathbf{b}_m + \tau))}{\sum_{j=1}^{J}\exp(\mathbf{x}(\mathbf{b}_j + \tau))}
\end{aligned} \tag{7}$$

Deshalb wird oft der Constraint $\mathbf{b}_1 = 0$ verwendet. Andere Constraints sind ebenfalls möglich.

Wenn $\mathbf{b}_1 = 0$ gilt $\exp(\mathbf{x}\mathbf{b}_1) = \exp(\mathbf{x}0) = 1$. Daher wird das MNLM oft wie folgt geschrieben:

$$\begin{aligned}
\Pr(y = 1|\mathbf{x}) &= \frac{1}{1 + \sum_{j=2}^{J}\exp(\mathbf{x}\mathbf{b}_j)} \\
\Pr(y = m|\mathbf{x}) &= \frac{\exp(\mathbf{x}\mathbf{b}_m)}{1 + \sum_{j=2}^{J}\exp(\mathbf{x}\mathbf{b}_j)} \quad \text{for } m > 1
\end{aligned} \tag{8}$$

4 Das MNLM als Discrete-Choice-Modell

Nach der Rational-Choice-Theorie wählt ein Individuum stets die Handlung, die ihm den höchsten Nutzen verspricht.

Wenn u_1 der Nutzen der Handlung 1 ist, und u_2 der Nutzen der Handlung 2, so wirde Handlung 1 gewählt, wenn

$$u_1 > u_2 \tag{9}$$

Die Einschätzung der Individuen über u setzen sich zusammen aus dem wahren Nutzen der Handlung μ_m und einem Fehler ϵ_m:

$$u_m = \mu_m + \epsilon_m \tag{10}$$

Demnach ist

$$\begin{aligned}\Pr(y = 1) &= \Pr(u_1 > u_2)\\ &= \Pr(\mu_1 + \epsilon_1 > \mu_2 + \epsilon_2)\\ &= \Pr(\epsilon_1 - \epsilon_2 > \mu_2 - \mu_1)\end{aligned}$$

bzw., bei J Alternativen

$$\Pr(y = m) = \Pr(u_m > u_j \text{ für alle } j \neq m) \tag{11}$$

Die spezifische Form eines Discrete-Choice-Modells hängt von der Annahme über die Verteilung der Fehler und dem Modell für μ_m ab. Bei $\mu_m = \mathbf{xb}_m$ und einer "type I extrem-value" Verteilung resultiert das MNLM.

5 ML-Estimation

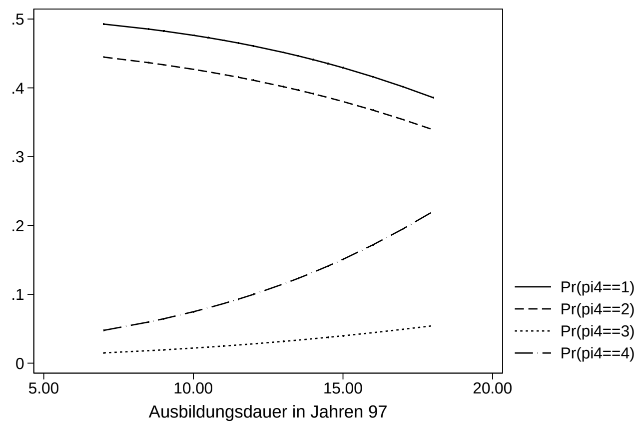
```
capture program drop mymlogit4
program define mymlogit4
    version 7.0
    args lnf theta1 theta2 theta3
    quietly replace `lnf' =  /*
    */ ln(1/(1 + exp(`theta1') + exp(`theta2') + exp(`theta3'))) /*
    */ if $ML_y1 == 1
    quietly replace `lnf' =  /*
    */ ln(exp(`theta1')/(1 + exp(`theta1') + exp(`theta2') + exp(`theta3'))) /*
    */ if $ML_y1 == 2
    quietly replace `lnf' =  /*
    */ ln(exp(`theta2')/(1 + exp(`theta1') + exp(`theta2') + exp(`theta3'))) /*
    */ if $ML_y1 == 3
    quietly replace `lnf' =  /*
    */ ln(exp(`theta3')/(1 + exp(`theta1') + exp(`theta2') + exp(`theta3'))) /*
    */ if $ML_y1 == 4
end

ml model lf mymlogit4 (CDU:pi=age) (FDP:pi=age) (B90:pi=age)
ml maximize
```

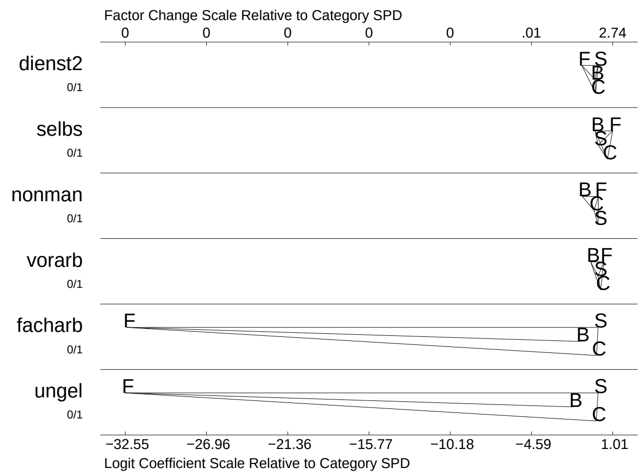

6 Interpretation

```
. gen pi4 = 1 if np9402==1 & np9402<=5
. replace pi4 = 2 if inlist(np9402,2,3) & np9402<=5
. replace pi4 = 3 if np9402==4 & np9402<=5
. replace pi4 = 4 if np9402==5 & np9402<=5
. lab val pi4 pi4
. lab def pi4 1 "SPD" 2 "CDU" 3 "FDP" 4 "B90"
. mlogit pi4 bdauer, base(4) nolog
```

```
. predict PSPD PCDU  PFDP PB90
. line P* bdauer, sort
```




```
. gen age = 1997-gebjahr
. gen men = sex==1
. gen dienst1 = egph==1
. gen dienst2 = egph==2
. gen selbs = inlist(egph,4,5,6)
. gen nonman = inlist(egph,3,11)
. gen vorarb = egp==7
. gen facharb= egp==8
. gen ungel = inlist(egp,9,10)
. gen mis = egp==.
. mlogit pi4 men age bdauer dienst2-mis
. mlogplot dienst2 selbs nonman vorarb facharb ungel, ///
>   std(000000) prob(.05)
```



7 Conditional Logit Model (CLM)

Das MNLM ist ein Spezialfall des CLM. Das CLM basiert auf einer Datenstruktur anderen Datenstruktur als das MNLM:

```
. list pi4 men age bdauer in 1/5
```

	pi4	men	age	bdauer
1.	B90	1	39	18
2.	CDU	1	28	12
3.	SPD	0	60	12
4.	CDU	1	44	10
5.	SPD	0	77	10

```
. gen id = _n
. expand 4  // Create one record for each outcome
. by id, sort: gen alt =_n
. gen choice = alt==pi
. l id alt pi choice in 1/8, sep(4)
```

	id	alt	pi4	choice
1.	1	1	B90	0
2.	1	2	B90	0
3.	1	3	B90	0
4.	1	4	B90	1
5.	2	1	CDU	0
6.	2	2	CDU	1
7.	2	3	CDU	0
8.	2	4	CDU	0

Kodierung der unabhängigen Variablen:

```
. gen spd = alt==1
. gen cdu = alt==2
. gen fdp = alt==3
. gen b90 = alt==4
. foreach var1 of varlist spd cdu fdp b90 {
.   foreach var2 of varlist men age bdauer {
.     gen `var1'`var2' = `var1' * `var2'
.   }
. }
```

Mit dem Condidtional Effect Modell kann die Auszahl zwischen Alternativen modelliert werden, die innerhalb der einzelnen Gruppen unterschiedliche Eigenschaften aufweisen.

▷ Beispiel:

- Modalsplit: Verkehrsmittelwahl ist abhängig von den Fahrtzeiten der einzelnen Verkehrsmittels für jeden Befragten
- Paneldaten: Zusammenhang zwischen Wahlverhalten und Issueorientierung, wobei die Befragten ihre Issueorientierung im Zeitablauf ändern.