Multivariat analyse af kategoriske variable med tabelanalyse og -tabulate-

[2. udgave, 28.12. 2012]

Kim Mannemar Sønderskov

Institut for Statskundskab, Aarhus Universitet
ks@ps.au.dk

Multivariat analyse refererer til analysesituationer med to eller flere uafhængige variable og anvendes typisk når man ønsker at undersøge sammenhængen mellem to variable kontrolleret for én eller flere variable (såkaldte kontrolvariable eller 3. variable). Lineær regression er en oplagt multivariat analyseteknik såfremt den afhængige variabel antages at være intervalskalleret. Denne teknik beskrives i Kapitel 8 i Sønderskov (2011). For andre typer afhængige variable vil andre teknikker være oplagte; eksempelvis logistisk regression til binære afhængige variable, ordinal logistisk regression til ordinale afhængige variable og multinominal logistisk regression til afhængige variable på nominal skala (jf. Tabel 6.1 i Sønderskov (2011); Long & Freese (2006) beskriver disse teknikker). Tabelanalyse kan i visse tilfælde være et simpelt, men informativt alternativ til de nævnte logistiske regressionsteknikker; multivariat tabelanalyse med Stata er emnet i dette notat. Sønderskov (2013) beskriver, hvordan resltaterne fra sådanne analyser kan illustreres grafisk.

Multivariat tabelanalyse bygger på principperne fra bivariat tabelanalyse, hvor en sammenhæng mellem to variable undersøges ved at krydstabulere de to variable (se kapitel 6 i Sønderskov, 2011). I det multivariate tilfælde med én afhængig variabel, én primær uafhængig variabel og én eller flere kontrolvariable krydstabuleres de to variable for alle kombinationer af kontrolvariablene. Man undersøger med andre ord sammenhængen mellem den afhængige og den primære variable ved alle kombinationer af kontrolvariablene. Hvis sammenhængen mellem variablene er til stede i alle kombinationer af kontrolvariablene vil konklusionen være, at der er en sammenhæng, når der kontrolleres for kontrolvariablene.

I Stata kan multivariat tabelanalyse blandt andet foretages med tabulate, der er den samme kommando som kan anvendes i det bivariate tilfælde. Den eneste forskel er, at tabulate-kommandoen udføres for alle kombinationer af kontrolvariablene. Dette foretages nemmest med by-prefixet (jf. p. 32 og 36 i Sønderskov, 2011). I en situation, hvor man vil undersøge sammenhængen mellem den afhængige variabel *yvar* og den primære uafhængige variabel *xvar* kontrolleret for *k1* og *k2* vil man anvende følgende kommando, hvor *yvar* behandles som kolonnevariabel og der procentueres på kolonnevariablen:

```
. *Krydstabulering af yvar og xvar ved alle kombinationer af k1 og k2
. by k1 k2, sort: tabulate yvar xvar, column
```

Dette svarer til at indsætte *k1* og *k2* under Repeat command by groups på by/if/in-fanbladet i dialogboksen til tabulate.

Kommandoen kan skrives kortere med bysort eller endnu kortere med bys (jf. boks 2.4 i Sønderskov, 2011):

```
. *Som ovenfor men med bysort/bys
. bysort k1 k2: tabulate yvar xvar, column
. bys k1 k2: tabulate yvar xvar, column //som foregående men med bys
```

Som i det bivariate tilfælde kan man bestille eksempelvis gamma-koefficienten ved optionen gamma:

```
. *som ovenfor med rapportering af gamma
. bys k1 k2: tabulate yvar xvar, column gamma
```

I det følgende eksempel analyseres sammenhængen mellem holdning til immigranter og stemmeadfærd ved hjælp af survey data (Den danske valgundersøgelse fra 2005). Konkret undersøges det om respondenternes niveau af frygt for indvandring påvirkede om de stemte på partier fra 'rød' eller 'blå' blok ved det netop overståede folketingsvalg. Der kontrolleres for respondentens køn og den antages bla. at stemmeadfærd ikke påvirker ens syn på fremmede (holdningen til fremmede behandles altså som uafhængig variabel).

1a) Åbn datasættet 'Valg05Kap7.dta' der kan hentes fra http://ps.au.dk/soenderskov/stata/.

1b) Inspicer variablene blå, v186 og mand med kommandoen codebook (se evt. afsnit 2.2.2 i Sønderskov 2011).

blå er en dikotom variabel, der antager værdien 1 hvis respondenten stemte på blå blok og 0 hvis respondenten stemte på partier fra rød blok (blå er altså en dummyvariabel). mand er ligeledes dummyvariabel, hvor 1 indikerer, at respondenten er en mand. v186 er en kategorisk variabel med 6 kategorier, hvoraf fem af dem indikerer forskellige niveauer af frygt for indvandring, mens en kategori indikerer "ved ikke". I nærværende tilfælde frasorteres "ved ikke"-svar, hvorved variablen er ordinalskaleret, med faldende frygt for indvandring ved stigende værdier.

2a) Omkod v186 så værdien '8' omkodes til 'missing'. Overskriv v186 så value labels beholdes. Du kan bruge følgende kommando:

```
. recode v337 (8=.)
```

2b) Foretag en bivariat analyse af sammenhængen mellem v186 og blå, hvor de relative frekvenser og gamma udregnes (se evt. opgave 125b i Sønderskov 2011). Husk at procentuere ud fra den uafhængige variabel.

Gamma-koefficienten og fordelingen indikerer en negativ sammenhæng mellem fravær af frygt og tilbøjeligheden for at stemme på et parti fra blå blok.

3) Kontroller for køn (med variablen mand) ved hjælp af bysort-prefixet.

Når der kontrolleres for køn producers to krydstabeller – én for kvinder og én for mænd og det fremgår, at der er en negativ sammenhæng mellem v186 og blå og for både kvinder og mænd.

Dette indikerer, at sammenhængen fundet uden kontrol også er til stede efter kontrol for køn. Det fremgår dog også, at der er en lille forskel i sammenhængens styrke for kvinder og mænd, idet gamma-koefficienten er højere (i absolutte termer) for kvinder end for mænd. Tilsyneladende betyder synet på immigration marginalt mere for kvinders stemmeadfærd end for mænd. Dette er et tegn på interaktion; altså at en sammenhæng mellem to variable afhænger af værdien på en anden variabel (se Kapitel 10 i Sønderskov, 2011).

Løsninger

```
*Opgave 1
use http://ps.au.dk/uploads/media/Valg05Kap7.dta //1a,clear
codebook blå v186 mand
*Opgave 2
recode v337 (8=.) //a
tabulate blå v186 ,colum gamma //b
*Opgave 3
bys mand: tabulate blå v186 ,colum gamma
```

Referencer

Long, J.S. & Freese, J. (2006). *Regression models for categorical dependent variables using Stata*. College Station: Stata Press.

Sønderskov, K.M. (2011). *Stata – En praktisk introduktion*. København: Hans Reitzels.

Sønderskov, K.M. (2013). *Stolpediagrammer for kategoriske data med -catplot-*. Institut for Statskundskab. Aarhus Universitet.