Jakob Roland Munch
Økonomisk Institut
Københavns Universitet
26. september 2004

Diskrete valg-modeller

Dette notat giver en kort introduktion til diskrete valg-modeller. Formålet med notatet er at sætte læseren i stand til at forstå resultater udledt fra diskrete valg-modeller. De mere detaljerede økonometriske egenskaber er således ikke i fokus.

Der findes mange situationer hvor den hændelse, man ønsker at modellere er diskret i stedet for kontinuert. Vi kommer til at betragte tre eksempler; 1) medlemmer af arbejdsstyrken skal vælge om de vil være forsikret mod arbejdsløshed eller ej, 2) hvorvidt arbejdsløse har deltaget i aktivering eller ej og 3) tilflyttere til Hovedstadsregionen skal vælge hvilken kommune de ønsker at bosætte sig i. I de første to eksempler er der kun to valgmuligheder (binært valg), mens der i det tredje eksempel er adskillige valgmuligheder. Eksemplerne adskiller sig også ved at vi i de første to primært er interesseret i hvorledes individuelle karakteristika påvirker valget mellem at være forsikret eller ej, mens vi i det andet tilfælde er interesseret i hvorledes valg-specifikke karakteristika (dvs. kommunespecifikke karakteristika) påvirker kommunevalget.

1 Binære valg-modeller

Man kan betragte modeller der kan forklare en binær afhængig variabel (en 0/1 variabel) på samme måde som den simple OLS-regression, hvor man i stedet har en kontinuert afhængig variabel. I begge tilfælde ønsker man at bestemme sammenhængen mellem den afhængige variabel og en række forklarende variable. I binære valg-modeller søger man således at forklare sandsynligheden for at valg-variablen, Y, antager værdien 0 eller 1.

Betragt situationen hvor medlemmer af arbejdsstyrken enten er forsikret mod arbejdsløshed ($Y = 1$) eller ikke er forsikret mod arbejdsløshed ($Y = 0$). Det formodes at en række individuelle karakteristika, som f.eks. alder, uddannelse, ledighedsrisiko mv.

(indeholdt i vektoren x) påvirker dette valg således at sandsynligheden for at være forsikret hhv. ikke forsikret kan skrives

$$\begin{aligned} P(Y &= 1) = F(\beta' x), \\ P(Y &= 0) = 1 - F(\beta' x), \end{aligned}$$

hvor parametrene i β-vektoren angiver effekten af ændringer i x på sandsynligheden for at være forsikret. Der findes forskellige tilgange til hvordan højresiden - og dermed funktionen F - modelleres.

1.1 Den lineære sandsynlighedsmodel

Fra økonometri undervisningen på 2. år stiftes bekendtskab med den lineære sandsynlighedsmodel. Her specificeres en sædvanlig lineær regressionsmodel af formen

$$\begin{aligned} y &= E(y) + (y - E(y)) \\ &= \beta' x + \epsilon. \end{aligned}$$

Her er det udnyttet at $E(y) = F(\beta' x) = \beta' x$. Et problem med denne model er imidlertid, at der er intet der sikrer at forudsigelserne $E(y) = \beta' x$, der jo er sandsynligheder, tilhører intervallet $[0, 1]$.

1.2 Probit og logit

Derfor benyttes ofte normalfordelingen i specifikationen af F, dvs.

$$P(Y = 1) = F(\beta' x) = \Phi(\beta' x),$$

hvor Φ er fordelingsfunktionen for normalfordelingen. Her gælder der at $P(Y = 1) \to 1$ for $\beta' x \to \infty$ og $P(Y = 1) \to 0$ for $\beta' x \to -\infty$, så vi holder os i intervallet $[0, 1]$. Dette er den såkaldte probit model.

En anden ofte anvendt specifikation er logit modellen, hvor der i stedet for normal fordelingen benyttes den logistiske fordeling, dvs.

$$P(Y = 1) = F(\beta' x) = \frac{e^{\beta' x}}{1 + e^{\beta' x}}.$$

Probit og logit modellerne estimeres typisk ved maksimum likelihood estimation. Det skal bemærkes, at den marginale effekt af en ændring i en af de forklarende variable på sandsynligheden for at være forsikret er vanskeligere at fortolke end i den simple lineære regressionsmodel. Vi holder os dog ofte til kvalitative udsagn, således at det primært er fortegnet på parameterestimaterne vi er interesseret i. Dvs. hvis en forklarende variabel får estimeret en parameter med positivt fortegn (f.eks. den individuelle ledighedsrisiko), har denne variabel en positiv effekt på sandsynligheden for at være forsikret.

2 "Conditional logit"-modellen

I et andet eksempel skal vi som sagt betragte valget af kommune i Hovedstadsregionen. Da der er mange kommuner i Hovedstadsregionen er der i modsætning til den binære valgmodel mere end to valgmuligheder. I tilfældet med J valgmuligheder kan den afhængige variabel Y således antage værdierne $1, 2, .., J$. Modeller med adskillige valgmuligheder kan motiveres ud fra stokastisk nytteteori. Nytten for tilflytter i ved at vælge kommune k kan skrives

$$U_{ik} = \beta' z_k + \epsilon_{ik},$$

hvor fejlleddene ϵ_{ik} er identisk og uafhængig fordelt i henhold til Weibull fordelingen. Bemærk kun valgspecifikke karakteristika z_k har betydning for nytten. Hvis tilflytteren vælger kommune k antages det at denne kommune giver højest nytte, dvs. sandsynligheden for at kommune k vælges kan skrives

$$P(Y_i = k) = P(U_{ik} > U_{ij}, j = 1, 2, .., J, j \neq k)$$

hvilket ved brug af fordelingsantagelse for fejlleddene kan omskrives til

$$P(Y_i = k) = \frac{e^{\beta' z_k}}{\sum_{j=1}^{J} e^{\beta' z_j}}.$$

Denne såkaldte "conditional logit"-model estimeres ved maksimum likelihood estimation, og igen er vi primært interesseret i fortolkningen af fortegnene på estimerede parametre. Hvis en forklarende variabel får estimeret en parameter med positivt fortegn (f.eks. kommunernes serviceudgifter), har denne variabel en positiv effekt på sandsynligheden for at kommunen vælges.