Siamak Baradaran
sia@kth.se

Tillvägaghångssätt för skattning av körkortsmodell

1 Syfte med modellen

Syftet med denna forskning har varit att utveckla en beskrivande modell som kan hjälpa oss att förstå benägenheten hos individer att förvärva körkort ur ett beteendemässigt perspektiv. Modellen skall även möjliggöra analyser av förändrade policyer och vara känslig mot individernas förändrade förutsättningar med tiden och relevanta samhällstrender.

2 Modellens struktur

Körkortsmodellen är framställd med hjälp ett dynamiskt modellramverk för så kallade ”hazard”-modeller (även kallad ”durations”-modeller).
Modellen är tidsdynamisk det vill säga den tar hänsyn till förändringar i tiden. Denna tidsdynamik påverkar modellen på två olika sätt.

1. Som en första del är modellen känslig avseende på tidsmässiga trender i populationen, till exempel påverkar denna tidsdynamik modellen om andel personer som tar körkort i olika åldrar ändras succesivt inom populationen. Denna egenskap fångas i modellen av så kallade ”base-line hazard”.

2. Den andra tidsdynamiken påverkar modellen då de individuella attributen ändras med tiden till exempel genom förändrade inkomstförhållanden eller familjestorlek under de år individen observeras. Dessa brukar kallas för ”time varying” co-variater.

Matematisk kallas modell funktionen ”complementary log-log function” och ser ut enligt följande:

$$\log\left(-\log\left(1-\lambda\left(t_j|\mathbf{x}_i\right)\right)\right) = \alpha_j + \mathbf{x'_i\beta}$$

där:

$$\alpha_j = \log\left(-\log\left(1-\lambda_0\left(t_j\right)\right)\right)$$

och $\lambda_0\left(t_j\right)$ är så kallade base line hazard och avser populationstrender (enligt punkt 1 ovan) då alla covariater är lika med noll, det vill säga då vi enbart tittar på hur körkortsinnehavet förändras med tiden utan att ta hänsyn till individuella attribut.
$\mathbf{x_i}$ är en vektor av individuella attribut (co-variater). Dessa attribut kan vara tids oberoende såsom kön, eller tidsberoende (time-varying) såsom ålder, inkomst, antal familjemedlemmar, etc.

$$\mathbf{x_i\beta} = x_i^1\beta_1 + x_i^2\beta_2 + \cdots + x_i^n\beta_n$$

i och j avser tider då individer har observerats så x_i^1 avser attribut nummer 1 som har observerats under tidsperiod i. $\beta_1, \beta_2, \dots, \beta_n$ är parametrar vilka skattas av modellen för respektive attribut (x).

3 Indata och modellresultat

Modellen har skattats med hjälp av data från två unika datasätt används i detta projekt. Båda innehåller observationer för åren mellan 2003 och 2011. Första datasättet har hämtats från årliga individuella skattedeklarationer vilka existerar för samtliga vuxna individer i Sverige.

Datasättet innehåller individens socioekonomiska attribut (modellens co-variater). Det andra datasättet är från bilprovningen och innehåller information om individuella fordon och dess ägare på årsbasis. Datasättet har därefter kombinerats med individuella attribut från det första datasättet. På det sättet vi får information individernas socio-ekonomiska attribut och om de fordon de äger.

För att minska modellarbetets komplexitet har vi begränsat antal fordon som kan ägas av en och samma individ (eller hushåll) till ett maximum av två fordon.

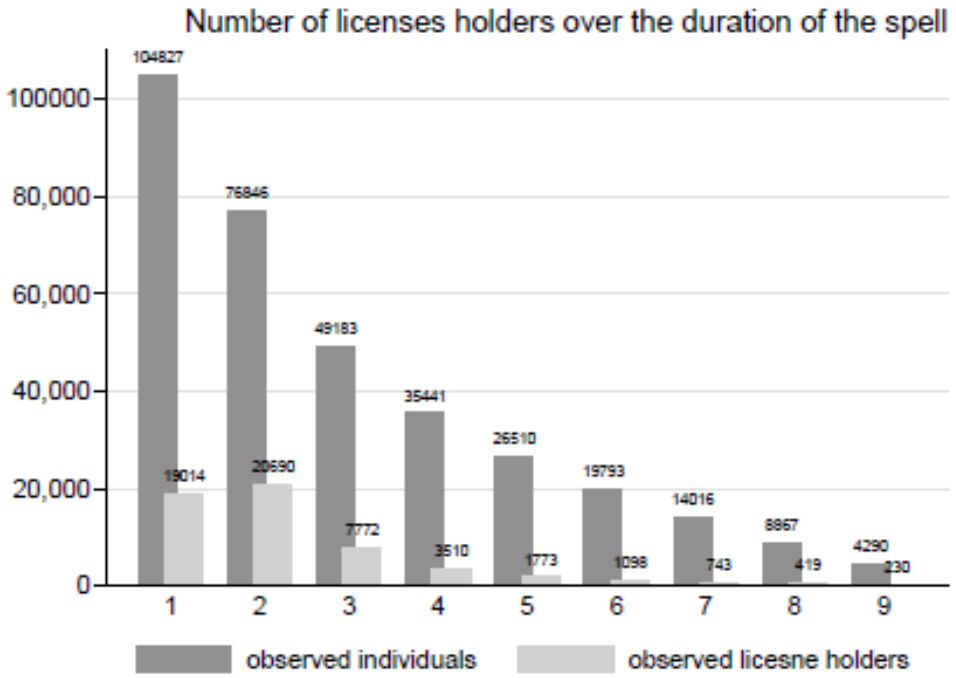
Datasättet består av ett stickprov som motsvarar 10% alla unika observerade vuxna i Sverige. Vi har helt enkelt tilldelat varje individ ett unikt ID-numer. Dessa ID är konsistent mellan åren för en och samma individ. Därefter har vi slumpmässigt valt 10 procent av individerna. Detta gör att de valda individerna finns i datasättet i olika många år. För en del finns observationer för alla år vi har data för och för somliga finns det för ett par år eller bara ett år.

Efter rensning av felaktiga observationer och observationer med tomma attribut vi lyckade samla nästan 340 000 observationer gjorda bland drygt 117 000 unika individer. I datamaterialet tillkommer cirka 11.000 nya individer årligen observeras in stavnings årligen. Figur 3 visar att nästan 20% av de individer som har observerats mellan 2003-2011 skaffade körkort samma år som de fyllde 18 år. Vidare ser vi att flest individer skaffar körkort ett år efter, det vill säga då de är 19 och att andelen som skaffar körkort efter det minskar succesivt men snabbt.

Som det framkommer innehåller datasättet många observationer och i och detta datasätt är inget undantag vad gäller problem. Materialet innehåller felaktigheter vilka vi har försökt åtgärda på olika sätt.

Det är dock inget bra ide att beskriva problemen här (de är hel enkelt många). Vi har för vårt ändamål skrivit ett så kallat script i programvaran STATA och med hjälp av det läser vi in de olika data sätten, sätter ihop individ och fordonsattributen, rensar materialet för olika problem och slutligen förbereder materialets format för modellen. I detta script har

vi kommenterat vad som görs så om man vill upprepa modellarbetet eller vill skatta om modellen kan man använda sig av scriptet. Observera dock att skriptet är drygt 2000 rader lång och på grund av det skickas scriptet separat och i digitalt format.



Tabellen nedan redogör för skattade parametrar för tre modeller med logistisk, kvadratisk och kubisk fördelningsantaganden. Modellstatistiken (grå område i figuren) visar att modellen med logistisk antagande är att föredra.

	logisitc	quadratic	cubic
age	-0.363***	-0.107***	-0.0118***
gender	-0.200***	-0.205***	-0.205***
employment status	0.676***	0.700***	0.696***
student	-0.086***	-0.225***	-0.203***
num. of children	0.278***	0.152***	0.160***
ln(parenets income)	0.539***	0.218***	0.230***
vehicles access	0.467***	0.450***	0.451***
ln(population)	-0.176***	-0.189***	-0.189***
ln(time)	0.535***		
$time^2$		-0.023***	
$time^3$			-0.004***
num. of observations	117755	117755	117755
$log-likeliehood$	-62075.5	-62216.3	-62175.4
degrees of freedom	9	9	9
AIC	124169.1	124450.6	124368.8
BIC	124256.2	124537.7	124455.9

Table 1: Estimated model parameters for logistic, quadratic and cubic models.

Kön (gender) är en dummy variabel där men representeras med värdet 0 och kvinnor med 1. Sama gäller variabeln student (1= student och 0 annars).

Egen inkomst visade sig vara en besvärlig variabel att använda då i de unga åren, inkomsten är nästan i perfekt samvariation med ålder. Vi valde att istället använda föräldrarnas inkomst. Populationen har använts som proxy for tillgänglighet. Hypotesen är att mindre städer tenderar att ha sämre kollektivtrafik och därmed större tendens hus befolkningen att skaffa körkort medan de som bor i större städer åtnjuter bättre kollektiv-trafik (och högre bilinnehavskostnader inklusive parkeringsavgifter) vilket bör minska deras benägenhet at skaffa bil och därmed även körkort.

Dessa skattade parametervärde kan användas för at skatta risken för varje enskild individ att skaffa körkort.

4 Skattning av individers benägenhet att skaffa körkort för prognos året¨

För prognosåret kan motsvarande benägenhet skattas om vi känner till individernas attribut för prognosåret. Emellertid finns inte information om variablerna vi är intresserade av för framtiden och någon slags hybrid data måste uppskattas eller simuleras, givet historisk individ-data. Simulering av data avseende framtida individuella attribut har dock inte ingått i detta projekt.

Vi kan som exempel på sådana simuleringsmodeller nämna simuleringsproceduren ”Survsim”[1] som finns tillgänglig för och modelleringsverktyget STATA.

[1] *Simulating complex survival data M.J. Crowther, P.C. Lambert, The Stata Journal (2012) 12, Number 4, pp. 674–687, http://www.stata-journal.com/sjpdf.html?articlenum=st0275*