Stockholms universitet

Användning av logistisk regression vid bedömning av chansen att lyckas vid uppkörningen av B-körkort

Tomas Hjert

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2015:7
Matematisk statistik
Juni 2015

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2015:7**
http://www.math.su.se/matstat

Användning av logistisk regression vid bedömning av chansen att lyckas vid uppkörningen av B-körkort

Tomas Hjert*

Juni 2015

Sammanfattning

Syftet med denna uppsats är att ta reda på hur tid, plats och personliga egenskaper påverkar sannolikheten att lyckas med uppkörning av B-körkort. För att skatta sannolikheten anpassas fyra preliminära logistiska regressionsmodeller, med godkänd eller underkänd som binär responsvariabel. Modellernas prediktionskraft analyseras och tolkas med ett flertal valideringstest. Samtliga modeller presterar relativt jämnt i prediktering, dessvärre inte fullt tillfredsställande. Vi får däremot en ganska tydlig bild av vilka variabler som har klinisk betydelse och i vilken utsträckning, där den mest betydelsefulla faktorn är huruvida man kör upp genom trafikskola eller som privatist.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: Tomashjert@hotmail.com. Handledare: Martin Sköld och Tom Britton.

Abstract

The purpose of this paper is to find out how the time, place and personal characteristics affect the likelihood of success with the driving test of catagory 'B'. To estimate the probability four preliminary logistic regression models are created, with passed or failed as binary response variable. The models predicting power are analyzed and interpreted by numerous validation tests. All models are performing relatively even in the prediction, unfortunately not fully satisfying. We get, however, a fairly clear picture of which variables have clinical significance and to what degree, were the most important factor is whether you take the test through the driving school or as a private.

Förord

Jag vill börja med att rikta ett stort tack till mina två handledare Tom Britton och Martin Sköld som bidragit med råd och stort engagemang under arbetets gång. Stort tack till trafikverket för datamaterialet.

Ett stort tack går även ut till Mikael Widlund och Josefin Andersson Senko för ovärdelig hjälp och rådgivning.

Detta är ett examensarbete motsvarande 15 högskolepoäng i matematisk-statistik. Arbetet är skrivet vid matematiska institutionen på Stockholms Universitet.

Innehåll

1 Introduktion

Att ta körkort är många ungdomars dröm och därför också ett vanligt samtalsämne bland såväl unga som gamla. Behov av körkort varierar mycket beroende på person, var i landet man bor, vissa behöver körkort för att underlätta sin dagliga livsföring medan andra endast tar det för nöjes skull. Gemensamt för alla är att det kostar pengar och tar tid vilket ger den logiska följden att det uppstår åtskilliga teorier, antaganden och hypoteser om hur man på bästa och enklaste sätt kan ta körkort.

Högst andel lyckade uppkörningar i landet har Örnsköldsvik, lägsta siffran återfinns i Stockholm, Sollentuna. Skillnaden är så stor som 29 procentenheter (42% resp 71%), finns det någon eller några faktorer som kan förklara detta? Är någon faktor mer dominant?

1.1 Syfte och metodik

Syftet med den här uppsatsen är alltså att undersöka vad som påverkar sannolikheten att lyckas med en uppkörning av B-körkort, och i vilken utsträckning. En uppsättning beskrivande faktorer används för att med logistisk regression skapa tre huvudsakliga modeller med hjälp av tre olika metoder. En modell kommer tas fram genom minimering av AIC, en andra genom minimering av BIC samt den tredje genom en metod kallad *purposeful selection*

1.2 Beskrivning av data

All data som används är mottaget av trafikverket och består av totalt 246 452 uppkörningar under kalenderåret 2014. Den data som används är vald utifrån det jag tyckt kan vara intressant och väsentligt för skattning av chansen att lyckas med en uppkörning av B-körkort. Variablerna som är med i urvalet till framtida modeller visas i Tabell 1, där de är kort presenterade. Det som kan vara värt att notera är att de kontinuerliga variabler som angivits snarare är diskreta med cirka 30-100 olika nivåer.

Tabell 1: Tabell över de variabler som är med i modellframtagningarna.

Variabel	Beskrivning	Typ	Nivåer
Säsong	Vinter = dec-feb Vår = mar-maj Sommar = jun-aug Höst = sep-nov	Nominal	Vinter Vår Sommar Höst
Dag	Vilken dag i veckan man kör upp	Nominal	Måndag-Söndag
Kön	Man eller kvinna	Nominal	0 = Kvinna 1 = Man
Utbildare	Om man kört upp genom trafikskola eller privat.	Nominal	0 = Privatist 1 = Trafikskola
Ålder	Ålder i heltal på personen som kör upp	Kontinuerlig	17-94
Starttid	Tid på dygnet man kör upp, mätt i timmar efter midnatt	Kontinuerlig	6.00-19.25
Latitud	Vilken latitud kontoret befinner sig på, högre värde = längre norrut	Kontinuerlig	55.60-65.58

2 Teori

Här presenteras en del teori som behövs och som använts i framtagandet av modellerna. Majoriteten av teorin är tagen ifrån Agresti (2002) *Categorical Data Analysis*, samt Hosmer och Lemeshow (2013) *Applied Logistic Regression.*

2.1 Varför logistisk regression?

Om vi låter Y vara en binär responsvariabel, där den antar antingen värdet 0 eller 1 för varje utfall. Väntevärdet för Y är väntevärdet för ett lyckat utfall och noteras $E(Y) = P(Y = 1|\mathbf{x})$. Vi definierar $P(Y = 1|\mathbf{x})$ som $\pi(\mathbf{x})$ för att visa på beroendet av värden $\mathbf{x} = (x_1, ..., x_p)$ från p prediktorer.

Den linjära sannolikhetsmodellen, $\pi(\mathbf{x}) = \alpha + \boldsymbol{\beta x}$, har sina begränsingar för att skatta sannolikheten. Sannolikheter påträffas endast mellan 0 och 1, medan linjära funktioner inte har några begränsningar för vilka värden de kan anta. Detta leder till att vissa skattningar i den linjära sannolikhetsmo-

dellen, $\hat{\pi}(\mathbf{x})$, riskerar att hamna utanför [0,1].

För binär data antas vanligtvis ett olinjärt samband mellan de oberoende variablerna $\mathbf{x}$ och $\pi(\mathbf{x})$, där en ändring i en oberoende variabel när $\pi(\mathbf{x})$ är nära 0 eller 1 oftast har mindre påverkan än då $\pi(\mathbf{x})$ är nära 0.5. Vanligt förekommande är så kallade S-formade kurvor och den viktigaste kurvan av den formen har funktionen

$$\pi(\mathbf{x}) = \frac{\exp(\alpha + \boldsymbol{\beta x})}{1 + \exp(\alpha + \boldsymbol{\beta x})}. \tag{1}$$

Denna benäms som den *logistiska regressionsmodellen* och har egenskapen att då $\boldsymbol{\beta x} \to \infty$ går $\pi(\mathbf{x}) \to 1$, och då $\boldsymbol{\beta x} \to -\infty$ går $\pi(\mathbf{x}) \to 0$.

Vi definierar oddsen som

$$\Omega = \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} = \exp(\alpha + \boldsymbol{\beta x}), \tag{2}$$

där oddsen är sannolikheten för positivt utfall dividerat med sannolikheten för negativt utfall. Logaritmering av ekvation (2) ger det linjära sambandet

$$\log \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} = \alpha + \boldsymbol{\beta x} = logit[\pi(\mathbf{x})]. \tag{3}$$

Detta ger den sökta länkfunktionen som är en logaritmering av oddsen, även kallad *logiten*. Logistiska regressionsmodeller, eller *logit modeller*, har egenskapen att *logiten* kan anta vilket reellt tal som helst, medan sannolikheten $\pi(\mathbf{x})$ alltid hamnar inom [0,1], vilket inte var fallet i den linjära sannolikhetsmodellen.

2.2 Odds-kvot

Från ekvation (2) ser vi att då sannolikheten ligger mellan [0,1] kommer oddset alltid vara ickenegativt, och då $\Omega > 1$ är sannolikheten för positivt utfall större än sannolikheten för negativt utfall.
Om vi noterar oddsen för lyckad uppkörning för män vid vissa förutsättningar som Ω_M, och oddsen för lyckad uppkörning för kvinnor under samma förutsättningar som Ω_K, då betecknas odds-kvoten som

$$\theta = \frac{\Omega_M}{\Omega_K} = \frac{\pi_M/(1 - \pi_M)}{\pi_K(1 - \pi_K)}, \tag{4}$$

där π_M är sannolikheten för lyckad uppkörning för män och π_K sannolikheten för kvinnor. Om $\theta > 1$ innebär det att sannolikheten för män att lyckas är större än för kvinnor. Om $\theta < 1$ har kvinnor större sannolikhet än männen, och om $\theta = 1$ är utfallet oberoende av kön.

2.3 Tolkning av parametrar

Studerar vi ekvation (1) noterar vi att tecknet på β_i bestämmer huruvida $\pi(\mathbf{x})$ ökar eller minskar i takt med x_i, om vi håller resterande x_j fixerade. Hastigheten för stigningen eller minskningen ökar i takt med att $|\beta_i|$ ökar, och då $\beta_i = 0$ är responsvariabeln oberoende av x_i.
Ekvation (2) visar att oddsen är en exponentiell funktion av $\mathbf{x}$, och kan användas för att tolka magnituden av β. En ökning av storlek 1 för x_i ger

$$\exp(\alpha + \beta_1 x_1 + ... + \beta_i(x_i + 1) + ... + \beta_p x_p) = \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} e^{\beta_i}. \qquad (5)$$

Ekvation (5) visar att en ökning av x_i med 1 ger en multiplikativ ökning i oddsen med e^{β_i}.

2.4 Purposeful selection

En möjlighet att ta fram en modell är att använda sig av *Purposeful selection* (Hosmer och Lemeshow (2013)). Det är en metod som består utav följande sju steg:

Steg 1: Purposeful selection börjar med att enskilt undersöka de olika oberoende variablerna med hjälp av en envariabels logistisk regressionsmodell. Här är man inte lika strikt i sin bedömning av vilket p-värde som krävs för att inkludera variabler, utan brukar använda så höga gränser som 0.2 eller 0.25. Detta på grund av att standardgränsen 0.05 oftast missar variabler som kan ha betydelse i en slutgiltig modell tillsammans med övriga variabler.

Steg 2: Inkludera nu samtliga variabler valda i steg 1 i en multivariat modell. Observera samtliga kovariaters p-värde från Wald-statistikan. De variabler som inte håller de vanliga signifikansnivåerna ska exkluderas ur modellen, varefter den nya, mindre modellen, ska jämföras med den gamla med hjälp av ett partiellt likelihood ratio test.

Steg 3: Jämför nu de skattade koefficienterna i den nya mindre modellen med den gamla. Speciellt intresserade är vi av de variabler med skillnader överstigande $\Delta\hat{\beta} > 20\%$. Detta då det indikerar att de variabler som uteslutits påverkar andra variablers skattningar. Om så är fallet ska dessa variabler in i modellen igen. Detta repeteras tills vi anser oss ha en modell med alla betydelsefulla variabler inkluderade.

Steg 4: Lägg nu till alla de variabler som inte blev valda i steg 1 till modellen, en i taget, och kontrollera dess signifikans. Detta är ett viktigt steg för att hitta och identifiera variabler som i sig inte är signifikanta men bidrar i samband med andra variabler. Modellen som kvarstår kallas *preliminära huvudeffektsmodellen.*

Steg 5: Vi tittar nu närmare på alla variabler i modellen. Nivåerna i de kategoriska variablerna ska vara rimliga och varje kontinuerlig variabel i modellen ska vara linjär i logiten. Modellen efter detta steg kallas *huvudeffektsmodellen*

Steg 6: Betrakta nu de samspel som kan vara aktuella. Ett samspel mellan variabler innebär att effekten av en variabel inte är konstant över olika nivåer av den andra variabeln. Huruvida ett samspel ska läggas till eller inte ska inte bara motiveras statistiskt utan även ur ett realistiskt perspektiv. Vi lägger till varje samspelsmöjlighet, en och en, och tittar enbart på den statistiska signifikansen på samspelet. Samspelet inkluderas om den klarar den vanliga 0.05 eller t.o.m. 0.01 gränsen. Efter vi bedömt vilka samspel som ska inkluderas, adderas samtliga till vår modell i steg 5. Vi implementerar åter igen steg 2, men anser nu att huvudeffekterna är låsta och inte kan tas bort. Modellen i slutet av detta steg kallas *preliminär slutmodell.*

Steg 7: Innan någon modell kan är anses slutgiltig, måste den genomgå ett flertal goodness of fit tester (se avsnitt **3**). Notera att detta steg även gäller för alla modeller och inte enbart de som framtagits med purposeful selection.

2.5 AIC och BIC

Ett av de mer välkända testen för att jämföra modeller är med hjäp av *Akaike Information Criterion*, AIC. Det är ett test för att jämföra modeller från samma dataset och säger alltså ingenting om hur bra modellen i helhet. Den är definierad som

$$AIC = 2p - 2ln(L) = -2(ln(L) - p),$$

där p är antalet parametrar och L är maximala Likelihoodfunktionen för modellen. Givet ett antal modeller så är den som föredras den med **lägst** AIC. AIC belönar alltså för höga likelihoodvärden och straffar för fler parametrar.

Ett liknande sätt att jämföra modeller är användandet av *Bayesian Information Criterion*, BIC. Det är snarlikt AIC med skillnaden att det tar hänsyn till antalet observationer. Det definieras som

$$BIC = \ln(n)p - 2\ln(L).$$

Skillnaden är att 2an framför antalet parametrar i AIC har bytts ut mot $\ln(n)$. Det innebär att BIC straffar modeller hårdare för fler parametrar, då $n > 7$, och uppmuntrar därmed till enklare modeller.

För den som vill läsa mer om olika informations-kriterium rekommenderas Dziak (2012) *Sensitivity and Specificity of Information Criteria.*

3 Goodness of fit

Två vanliga tester för hur väl en modell presterar är att undersöka modellens devians, eller Pearsons statistika. Dessa fungerar inte som teststatistikor för data som är helt ogrupperad (Hallet (1999) s.12-13). Eftersom vi har nästintill ogrupperad data använder vi inte dessa.

3.1 Hosmer Lemeshow test

Om vi har n skattade sannolikheter för n olika händelser, rangordnar vi dom i storleksordning där den första kolumnen är den lägsta skattningen och den n:te kolumnen är den högst skattade sannolikheten. Hosmer-Lemeshow är ett test som delar upp dessa skattade sannolikheter i, vanligtvis, $g = 10$ grupper, där den första gruppen består av de lägsta tio procenten av skattade sannolikheter, och den sista gruppen består av de högst skattade sannolikheterna. Pearson chi-squared statistikan används sedan för jämförelse mellan observerade och skattade värden och resultatet blir en χ^2-fördelad statistika, $\widehat{C}$, med (g-2) frihetsgrader. Om $\widehat{C}$ är signifikant visar det på att modellen ej passar väl till underliggande data. Problemet är att χ^2-värdet ökar med antalet observationer, vilket renderar i att alla modeller förkastas genom för låga p-värden då mängden data är tillräckligt stort (Hosmer Lemeshow (2013) s.168).

Vi använder detta test trots vår mängd data, men då främst för jämförelse mellan modellers χ^2värden.

3.2 Receiver Operating Curve, ROC

Summerar man resultatet av en logistisk regression med hjälp av en 2x2 klassificerings-tabell, noterar man utfallet av den binära responsvariabeln, givet skattningen av respektive utfall. Skattningen ansätts som antingen 1 eller 0, lyckad eller misslyckad, beroende på skattningen av sannolikheten på utfallet och en "cutpoint"c. Om händelse i har skattad sannolikhet $\widehat{\pi}_i$, blir

$\widehat{y}_i$=1 om $\widehat{\pi}_i > c$, och $\widehat{y}_i = 0$ om $\widehat{\pi}_i < c$. Man summerar prediktionskraften genom definitionerna (Agresti (2002) s.228)

$$\text{sensitivitet} = \text{P}(\hat{y} = 1|y = 1) \text{ and specificitet} = \text{P}(\hat{y} = 0|y = 0).$$

En Receiving Operating Curve, ROC, är en plot som plottar sensitivitet mot (1-specificitet) för alla möjliga "cutpoints". Kurvan som bildas är oftast konkav och binder ihop punkterna (0,0) och (1,1). Arean som bildas under kurvan återspeglar prediktionskraften där större area återspeglar högre prediktionskraft. Arean är även identisk mot ett annat mätvärde av prediktionskraft, *concordance index*. En area på 0.5 återspeglas av en rät linje i plotten och säger att modellen inte är bättre än slumpmässiga gissningar (Agresti (2002) s.229). Riktlinjer av hur man ska tolka arean under ROC ges av (Hosmer och Lemeshow (2013) s.177):

$$Om \begin{cases} AUC = 0.5, & \text{Inte bättre än slumpmässiga gissningar} \\ 0.5 < AUC < 0.7, & \text{Anses som dåligt, inte mycket bättre än gissningar} \\ 0.7 < AUC < 0.8, & \text{Anses acceptabelt} \\ 0.8 < AUC < 0.9, & \text{Anses mycket bra} \\ AUC \geq 0.9. & \text{Anses utomordentligt bra} \end{cases} \tag{6}$$

Exempel på hur en ROC kan se ut visas i Figur 1.

3.3 Generaliserande R^2

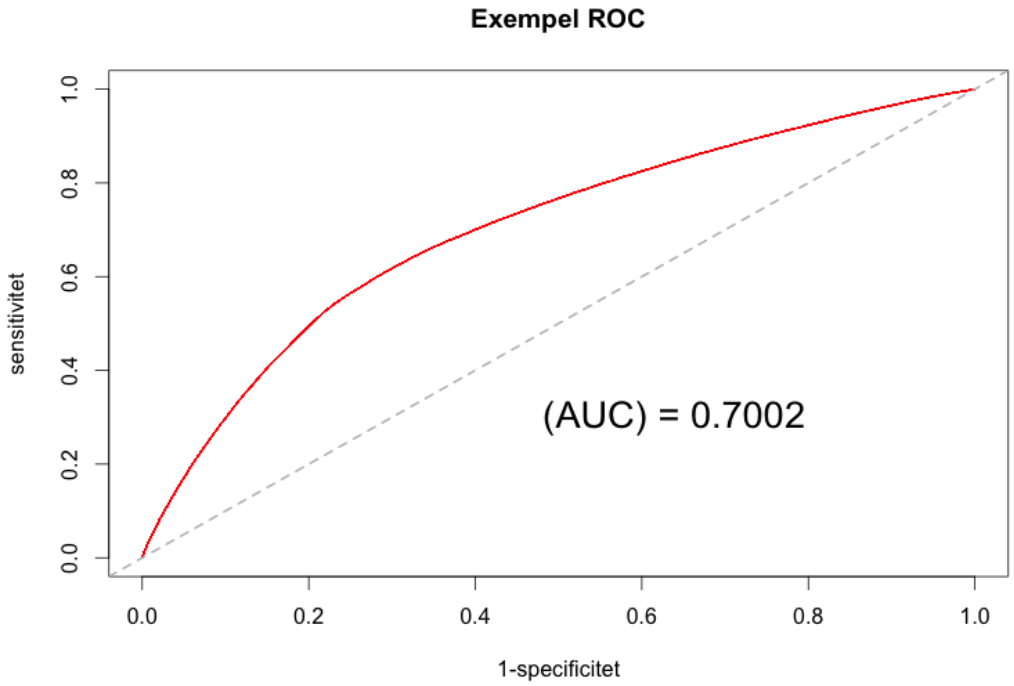
I linjär regression stöter man ofta på uttrycket R^2 som förklarar hur stor del av variansen som förklaras av de oberoende variablerna under linjära förhållanden. Formeln för R^2 ges av

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2},$$

där y_i är observerade värdet, $\hat{y}_i$ skattade värdet och $\bar{y}$ medelvärdet av observationerna. I logistisk regression är den inte lika användbar men många har försökt skapa en R^2 anpassad för lite mer generella modeller där maximum likelihoodanpassning använts. Cox and Snell med flera föreslog följande typ av generaliserande R^2 (Hallet (1999) s. 36)

$$R_g^2 = 1 - [\hat{L}_c/\hat{L}_0]^{2/n},$$

där $\hat{L}_c$ representerar log likelihooden uträknad i $\hat{\boldsymbol{\beta}}$ och $\hat{L}_0$ representerar log likelihooden uträknad endast för interceptmodellen. Nagelkerke (1991) föreslog



Figur 1: Exempel på en ROC med inkluderande AUC, *concordance index*.

en modifiering så att R_g^2, likt den linjära diton, kan ligga mellan (0,1) genom följande ändring

$$\bar{R}_g = R_g^2 / \max(R_g^2),$$

där $\max(R_g^2) = 1 - (\hat{L}_0)^{2/n}$.

Dessa värden är dock ofta väldigt låga och ska ej användas i samma utsträckning som i enkel linjär regression. Den används här som komplement och jämförelseverktyg.

3.4 Skattade mot observerade

Ett mer visuellt hjälpmedel för att se hur modellen predikterar är att plotta skattade värden mot de observerade. Om modellen passar bra kommer denna plot visa en rät 45-gradig linje mellan punkterna (0,0) och (1,1). Då vi har många värden kommer vi inte skatta alla enskilt utan dela upp dem i 100 grupper rankat efter de skattade sannolikheterna. Fördelen med denna plot är att den ger ett enkelt visuellt hjälpmedel för att analysera hur modellen presterar över de olika sannolikhetsskattningarna.

4 Framtagning av modeller

Vi utgår nu från de sju variabler som givits i Tabell 1 för att få fram en modell som kan prediktera sannolikheten att lyckas vid uppkörning av B-körkort. Modellen ska prestera tillfredställande i prediktionssyfte men även vara enkel att tolka. Vi använder oss av tre olika metoder för att få fram tre preliminära slutmodeller, som vi sedan jämför och analyserar. Den första modellen tas fram med hjälp av *purposeful selection*, beskrivet i avsnitt **2.4**.

Steg 1: Som första steg undersöker vi enskilt varje kovariat och tittar på dess p-värde. Då mängden data är så stor, är det föga förvånande att samtliga blir signifikanta och det på flera decimalers noggrannhet.

Steg 2: Den multivariata modellen med samtliga variabler från steg 1 ses i Tabell 2. Vi ser att de enda kovariater med ett p-värde högre än 0.05 är de

Tabell 2: Resultatet av den multivariata modellen med samtliga kovariater från steg 1. Notera att den första nivån i varje kategori ingår i interceptet.

	Skattning	Std.Av	p-värde
(Intercept)	-2.54944	0.12993	$<$ 2e-16
SasongSommar	0.05572	0.01182	2.45e-06
SasongVår	0.03135	0.012	0.008967
SasongVinter	0.04797	0.01213	7.66e-05
Dagtisdag	0.01369	0.01363	0.31513
Dagonsdag	-0.01510	0.01366	0.26901
Dagtorsdag	-0.02005	0.01384	0.14746
Dagfredag	-0.09649	0.01483	7.74e-11
Daglördag	0.01973	0.03056	0.51846
Dagsöndag	-0.00036	0.15307	0.99810
KonMan	0.13658	0.00877	$<$ 2e-16
Alder	-0.03169	0.0005	$<$ 2e-16
UtbTrafikskola	1.23893	0.00919	$<$ 2e-16
Starttid	-0.01867	0.00180	$<$ 2e-16
Latitud	0.05203	0.00216	$<$ 2e-16

dummyvariabler som skiljer respektive dag mot måndag, med undantag fredag. Uteslutning av variabeln *Dag* från modellen ger ett G-värde på 68.2762, som med 6 graders frihet ger ett p-värde på ca $9 * 10^{-13}$. Och något ihopslag av dagar är svårt att motivera (utom kanske lördag och söndag). Då ingen

variabel har tagits bort så går vi direkt till steg 5.

Steg 5: För att undersöka linjäritetet i våra kontinuerliga variabler Ålder, Starttid och Latitud använder vi en *smoothed scatterplot* där vi plottar respektive variabel mot *logiten*, d.v.s log-oddsen. Kan sambandet där antydas som linjärt är det troligt att variabeln är linjär i *logiten*. Figur 3 i Appendix B visar resultaten av plottarna. Variabeln Ålder har en lite kraftigare sänkning första åren för att sedan se mer konstant nedåtgående ut. Ändringen är dock så pass liten att vi fortsätter anta linjäritet i logiten. Starttid har en svagt nedgående trend även den och ingenting säger oss att den inte är linjär i logiten. Latitud är lite mer svårtolkad, då den har en liten dipp i logiten mellan 58-60, vilket kan förklaras av de låga resultaten i Stockholmsområdena. Detta dock så pass lite att vi även här utgår från att Latitud är linjär i logiten och vi fortsätter med samtliga som sådana till steg 6. Vi kallar den modell från Tabell 2 för vår *huvudeffektsmodell*.
Steg 6: Vi undersöker eventuella samspel som kan finnas mellan de kvarvarande huvudeffekterna. Om vi utgår ifrån att det finns en samspelsmöjlighet mellan samtliga variabler, testar vi att lägga till de $\binom{7}{2} = 21$ olika möjligheterna en och en. Tabell 3 visar resultatet. På grund av det stora antalet observationer är vi hårda med vilken signifikansnivå vi använder och tittar endast på de samspel som har en signifikansnivå på under 0.001. Det är 6 st varav ingen jag kan fökasta med min ringa expertis inom ämnet och samtliga får därmed inkluderas i slutmodellen som visas i Tabell 7 i Appendix A.

4.1 Stepwise AIC och BIC

För att ta fram de modeller som har lägst AIC respektive BIC så används proceduren *step* i RStudio [8]. Proceduren börjar med en modell med samtliga variabler inkluderade exklusive samspel. Den modellerar sedan ett steg i taget där den värderar samtliga möjligheter, dvs antingen ta bort någon variabel eller lägga till ett samspel. Den väljer sedan det alternativ vars AIC/BIC som är lägst och fortsätter tills den modell man har är den modell med lägst AIC/BIC. Modellerna som fås fram på detta sätt visas i Tabell 8 och Tabell 9 i Appendix A.

Tittar man närmare på de samspel som finns med, främst de mest signifikanta som återfinns i *modellBIC*, ser man att deras bidrag till skattningarna förvisso är signifikanta men marginella. För att undersöka hur en modell utan dessa samspel presterar tas en ny modell fram för analys. Vi kallar den fjärde modellen för *modellTest* och visas i Tabell 10 i Appendix A.

Tabell 3: Möjliga samspel mellan huvudeffekterna om de skulle läggas till ensamma till vår huvudeffektsmodell från steg 5.

Samspel	Df	Deviance	AIC	LRT	p-värde
<none>		310789	310819		
Säsong:Dag	17	310766	310830	22.921	0.151832
Säsong:Kön	3	310765	310801	23.595	3.035e-05
Säsong:Ålder	3	310786	310822	2.944	0.400367
Säsong:Utbildare	3	310788	310824	1.047	0.789834
Säsong:Starttid	3	310783	310819	5.387	0.145549
Säsong:Latitud	3	310781	310817	7.620	0.054559
Dag:Kön	6	310783	310825	5.706	0.456865
Dag:Ålder	6	310778	310820	10.623	0.100742
Dag:Utbildare	6	310770	310812	18.722	0.004660
Dag:Starttid	6	310782	310824	6.669	0.352571
Dag:Latitud	6	310776	310818	12.666	0.048651
Kön:Alder	1	310564	310596	224.705	$< 2.2e-16$
Kön:Utbildare	1	310743	310775	45.649	1.415e-11
Kön:Starttid	1	310788	310820	0.404	0.525274
Kön:Latitud	1	310773	310805	15.520	8.163e-05
Ålder:Utbildare	1	310766	310798	23.261	1.414e-06
Ålder:Starttid	1	310779	310811	9.880	0.001671
Ålder:Latitud	1	310763	310795	25.405	4.647e-07
Utbildare:Starttid	1	310778	310810	10.794	0.001018
Utbildare:Latitud	1	310789	310821	0.025	0.874702
Starttid:Latitud	1	310786	310818	2.482	0.115120

5 Resultat och analys

Vi har nu fyra stycken preliminära modeller. Den som framtagits med hjälp av metoden purposeful selection (*modellPurp*), de två som tagits fram av RStudio som minimerat AIC (*modellAIC*), minimerat BIC (*modellBIC*), samt den förenklade *modellTest*. Vi testar nu hur de olika modellerna presterar på diverse tester för att mäta prediktskraft och styrka.

Hosmer Lemeshow test [10] genomförs på samtliga modeller enligt vad som beskrivits i avsnitt 4.1. Värderna som fås visas i Tabell 4.

Som väntat blir alla modellerna högst signifikanta på grund av det stora antalet observationer och ingen av modellerna håller måttet enligt detta test.

Tabell 4: Hosmer Lemeshows testresultat för de fyra alternativa modellerna.

Hosmer Lemeshow	ModellAIC	ModellBIC	ModellPurp	ModellTest
χ^2-värde (8 df)	207.4975	256.3163	223.5095	226.3022
p-värde	$<$ 2.2e-16	$<$ 2.2e-16	$<$ 2.2e-16	$<$ 2.2e-16

ModellAIC presterar till synes något bättre (mindre dåligt) än de övriga modellerna.

Receiver operating curve [9] anpassas till modellerna och arean under kurvorna, AUC eller *concordance index*, beräknas. Resultaten visas i Figur 6 och Figur 7 i Appendix B. Noterbart är att modellerna skiljer sig åt ytterst lite, och det är först på den tredje decimalen. Vi kan också se att samtliga modeller får ett AUC-värde strax under 0.7, vilket var gränsen för acceptabelt enligt Hosmer Lemeshow, se avsnitt 4.2.

Generaliserande R^2 [2] beräknas, både R_g^2 och Nagelkerges modifierade $\bar{R}_g^2$, och resulterade i följande värden.

Tabell 5: De generaliserade R^2 för de tre modellerna.

Typ av R^2	ModellAIC	ModellBIC	ModellPurp	ModellTest
Cox & Snells R_g^2	0.1175862	0.1169225	0.1173562	0.1156586
Nagelkerkes $\bar{R}_g^2$	0.1568831	0.1559975	0.1565762	0.1543113

Skattade mot observerade värden plottas mot varandra och resultaten för de fyra modellerna visas i Figur 4 och Figur 5 i Appendix B. Samtliga fyra modeller tenderar ha samma typ av utseende. Överlag ser det ut som att samtliga modeller har en acceptabel prediktionsförmåga.

Summering av resultaten tillsammans med AIC och BIC visas i Tabell 6.

Tabell 6 visar på att modellAIC presterar bäst på samtliga tester förutom BIC, där förklarligen modellBIC är bäst. Det är dock inte så oväntat då den har den mest komplexa uppbyggnaden och frågan blir då om komplexiteten överväger enkelheten i någon av de andra enklare modellerna. Alla tester ger dock väldigt snarlika värden på samtliga modeller.Efter inspektion av Figur 4 och 5 tillsammans med de värden testerna givit oss, blir slutsatsen att *modellTest*s extrema enkelhet kraftigt överväger de små fördelar de mer komplexa modellerna ger oss.

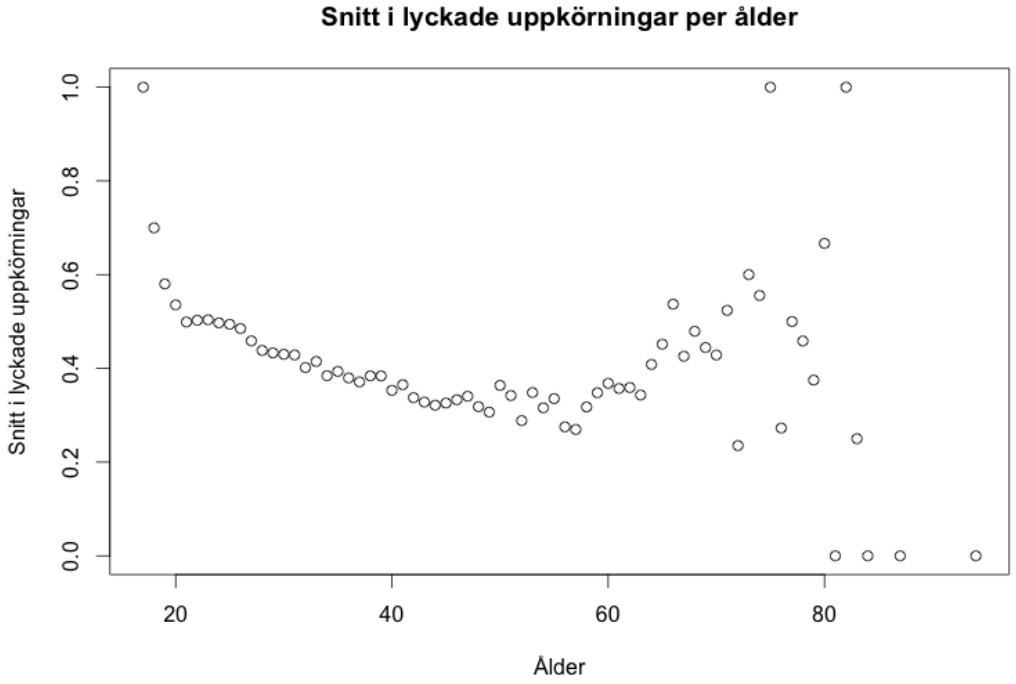
Tabell 6: Summering av de tre preliminära slutmodellerna.

	ModellAIC	ModellBIC	ModellPurp	ModellTest
Cox & Snells R_g^2	0.1175862	0.1169225	0.1173562	0.1156586
Nagelkerkes $\bar{R}_g^2$	0.1568831	0.1559975	0.1565762	0.1543113
χ^2-värde (8 df)	207.4975	256.3163	223.5095	226.3022
AUC	0.6989	0.6983	0.6987	0.6974
AIC	310429.5	310552.8	310457.7	310897.3
BIC	310856.5	310656.9	310697.3	310959.8

Med utgång från den enkla *modellTest* så räknar vi ut Odds-kvoten för de variabler som kvarstår i Tabell 11 i Appendix A. Detta ger att odds-kvoten ökar med 14.6% om du är man jämfört med om du är kvinna. För varje positivt steg i de kontinuerliga variablerna Ålder och Starttid, får vi en minskning i odds-kvoten med ca 3.1% resp 1.7%, medan för Latitud får vi för varje steg en skattad ökning i odds-kvoten med 5.3%. Den största påverkan på den skattade sannolikheten för lyckad uppkörning är dock huruvida man kör upp genom trafikskola eller som privatist. Där har den som kör upp genom trafikskola en skattad ökning i odds-kvoten på hela 247%. Allt detta kan illustreras till viss del av grafiska hjälpmedel där till exempel Figur 2 visar medvärdet av de som klarar uppkörningen per ålder. Figuren visar en ganska klar negativ trend fram till ungefär år 65, där det blir en kraftig variation som förklaras av väldigt få utfall (334 uppkörningar i åldrar över 65), vilket även är fallet i ålder 17 (1 uppkörning).

6 Slutsats

Med utgång från de resultat som visats så är det svårt att bedöma hur väl man kan lita på modellerna i prediktionssyfte. Samtliga värden tyder på en relativt dålig prediktionskraft men om det är på grund av det stora datamaterialet är svårt att avgöra. Med tanke på plottarna av observerade mot predikterade värden, som visade på att samtliga modellers skattande höll en acceptabel nivå, så förkastar vi inte möjligheten att använda modellerna i prediktionssyfte. Intressant är även hur pass liten skillnad det verkar vara i den mycket mer komplexa *modellAIC* jämfört med den väldigt enkla *modellTest*, och det gör att den senare rekommenderas oavsett syfte på grund av dess mycket simpla uppbyggnad.
Värt att notera är att variablerna *dag* och *säsong*, samt alla samspel som



Figur 2: Illustrerande bild över medelvärdet för lyckade uppkörningar för varje ålder.

ingår i *modellAIC*, endast ger marginella förbättringar jämfört med *modellTest*. Detta antyder att variablerna sinsemellan är oberoende.
Med utgång från *modellTest* ser man att för att maximera sina chanser att lyckas med sin uppkörning, ska man köra upp när man är ung, tidigt på dygnet, så långt upp i landet som möjligt och framför allt att köra upp genom trafikskola. Huruvuda man är man eller kvinna är dock svårt att påverka. Vilken dag i veckan samt vilken årstid uppkörningen sker har mindre betydelse.

7 Diskussion

Alla modeller är en förenkling av verkligheten, frågan är om den valda modellen är för mycket förenklad. Samtliga modeller visar på tveksam prediktionsförmåga, frågan är dock varför. Felaktigheter i kontinuerliga kovariater, fel i systematiska komponenter eller användandet av fel länk-funktion kan vara några av anledningarna (Hosmer och Lemeshow (2013) s.203). Eventuella transformationer av de kontinuerliga variablerna skulle kunna genomföras,

det skulle därför vara intressant att göra ytterligare studier med transformerade variabler för att se om detta förbättrar resultatet. Samspelseffekter visades endast påverka resultaten marginellt och om någon annan länk skulle förbättra resultatet vet jag inte, även detta skulle framtida forskning kunna besvara. Den kontinuerliga variabeln *ålder* visar på en negativ trend. I detta datamaterial ser man dock inte om det är första eller tionde gången en person kör upp. Det innebär alltså att personer som misslyckas flera gånger bidrar till datamaterialet med flera observationer.

Då åldern är konstant växande kan det tänkas att vissa årskullars skattning drabbas mer av de uppföljande misslyckandena än andra, till exempel skulle det kunna vara så att yngre personer kör upp flera gånger per år eller att äldre ålderskullar drabbas hårdare av återkommande misslyckande. Den negativa skattningen skulle kanske ändras om man endast skulle haft med förstagångsuppkörningar eller antalet uppkörningar som ytterligare variabler. Det skulle därför vara intressant att göra liknande studier på ett material där någon av dessa variabler ingår eller där vi delar upp data baserat på ålder för att finna skillnader i skattningarna. Det är tveksamt om det enbart är variabeln *latitud* som är orsak till dess positiva skattning, eller om det kan ligga andra orsaker bakom. Vi såg i Figur 3 i Appendix A att det var en dipp i logiten då latituden var runt 59, vilket är värdet för stockholmsområdet. Det skulle därför vara intressant att inkludera ett komplement eller alternativ till Latitud som exempelvis Tätort kontra Landsort, eller population för uppkörningsplatsen. Båda dessa variabler var aktuella för inkludering, svårigheter med klassificering gjorde att de hölls utanför denna studie. En annan anledning till varför det visade sig vara skillnad i landet kan vara motivation och tradition. Det finns ett större behov av bil och körkort i delar av landet där kollektivtrafiken är bristande och avstånden är större. Variabeln Tätort kontra Landsort motiveras återigen av detta.

Att datamaterialet är så pass omfattande försvårade då jag beaktade signifikansnivåer snarare än hur stor egentlig påverkan variabeln i fråga hade. Det var denna risk för överanpassning (over-fitting, (Hosmer och Lemeshow (2013) s.168)) som medförde att signifikanta variabler med marginell påverkan och samspel uteslöts ur modellen.

Ett alternativt utförande är att dela upp materialet i ett test-set, och ett validerings-set (Hosmer och Lemeshow (2013) s.202). Möjligen kan test-settet undvika att hitta triviala avvikelser samt användas på data som ej använts till modellanpassningen.

Ett alternativ till variabeln *säsong* är att jämföra vädret på uppkörningsdagen. Det är dock oklart hur tillgängligheten för sådan data ser ut. Det vore även intressant att se över andra typer av körkort för att se om dessa visar samma trender som vi ser för B-körkort. Det kan finnas flera variabler som kan bidra till en lyckad uppkörning, detta bör diskuteras inför en framtida studie.

Referenser

[1] Agresti, A. (2002). Categorical data analysis - 2nd ed. John Wiley & Sons, Inc.

[2] Beaujean, A. (2012). Bayloredpsych: R package for baylor university educational psychology quantitative courses. R package version 0.5, URL http://CRAN. R-project. org/package= BaylorEdPsych.

[3] Brett Presnell (2000) *An Introduction to Categorical Data Analysis Using R.* PDF File.

[4] Dziak, J. J., Coffman, D. L., Lanza, S. T., & Li, R. (2012). Sensitivity and specificity of information criteria. The Methodology Center and Department of Statistics, Penn State, The Pennsylvania State University.

[5] Hdett, D. C. (1999). Goodness of fit tests in logistic regression (Doctoral dissertation, Ph. D. thesis, University of Toronto).

[6] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.

[7] Moa Skagerlind (2015-05-08) *http://www.dn.se/ekonomi/har-lyckas-flest-att-kora-upp/*

[8] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[9] Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: visualizing classifier performance in R. Bioinformatics, 21(20), 3940-3941.

[10] Subhash R. Lele, Jonah L. Keim and Peter Solymos (2014). ResourceSelection: Resource Selection (Probability) Functions for Use-Availability Data. R package version 0.2-4. http://CRAN.R-project.org/package=ResourceSelection

[11] Thompson, L. A. (2009). S-PLUS (and R) manual to accompany Agresti's Categorical Data Analysis (2002).

[12] Venables, W. N., & Ripley, B. D. (2002). Modern applied statistics with S. Springer Science & Business Media.

Appendix

A Tabeller

Tabell 7: Preliminär slutmodell med samspelstermer, *modellPurp*

ModellPurp	Skattning	Std.Av	p-värde
(Intercept)	-4.66807	0.41377	< 2e-16
SasongSommar	0.01343	0.01755	0.444072
SasongVår	-0.03125	0.01800	0.082658
SasongVinter	0.00826	0.01815	0.649021
Dagtisdag	0.01454	0.01363	0.286199
Dagonsdag	-0.01460	0.01366	0.285194
Dagtorsdag	-0.01955	0.01385	0.158121
Dagfredag	-0.09536	0.01484	1.31e-10
Daglördag	0.02099	0.03056	0.492116
Dagsöndag	-0.00110	0.15298	0.994249
KonMan	0.56952	0.25841	0.027529
Alder	0.03631	0.01518	0.016793
UtbTrafikskola	1.19043	0.03036	< 2e-16
Starttid	-0.01857	0.00180	< 2e-16
Latitud	0.09294	0.00699	< 2e-16
SäsSommar:KonMan	0.07818	0.02375	0.000998
SäsVår:KonMan	0.11324	0.02414	2.72e-06
SäsVinter:KonMan	0.07186	0.02440	0.003240
KonMan:Alder	0.01643	0.00104	< 2e-16
KonMan:UtbTrafikskola	0.16388	0.01822	< 2e-16
KonMan:Latitud	-0.01655	0.00435	0.000144
Alder:UtbTrafikskola	-0.00126	0.00110	0.252329
Alder:Latitud	-0.00130	0.00025	3.94e-07

Tabell 8: Modell som minimerar AIC, ModellAIC

ModellAIC	Skattning	Std.Av	p-värde
(Intercept)	-4.7280	0.8905	1.10e-07
SasongSommar	1.0479	0.3512	0.002853
SasongVår	0.3613	0.3604	0.316129
SasongVinter	0.6349	0.3632	0.080437
Dagtisdag	-0.0918	0.3957	0.816503
Dagonsdag	0.0332	0.3951	0.932915
Dagtorsdag	0.8664	0.4034	0.031759
Dagfredag	-0.1846	0.4552	0.684939
Daglördag	4.1502	1.7110	0.015286
Dagsöndag	-0.9222	7.8562	0.906549
KonMan	0.6755	0.2629	0.010201
Alder	0.0368	0.0156	0.018315
UtbildareTrafikskola	1.7042	0.2711	3.29e-10
Starttid	-0.1055	0.0555	0.057535
Latitud	0.0952	0.0150	2.80e-10
KonMan:Alder	0.0166	0.0010	<2e-16
KonMan:UtbTrafikskola	0.1671	0.0182	< 2e-16
Alder:Latitud	-0.0013	0.0002	9.10e-08
SasongSommar:KonMan	0.0791	0.0237	0.000872
SasongVår:KonMan	0.1145	0.0241	2.10e-06
SasongVinter:KonMan	0.0735	0.0244	0.002607
KonMan:Latitud	-0.0184	0.0044	3.16e-05
UtbTrafikskola:Starttid	-0.0129	0.0037	0.000651
Dagtisdag:UtbTrafikskola	0.0460	0.0285	0.107347
Dagonsdag:UtbTrafikskola	0.0020	0.0284	0.941521
Dagtorsdag:UtbTrafikskola	0.0077	0.0289	0.789767
Dagfredag:UtbTrafikskola	-0.0524	0.0317	0.098401
Daglördag:UtbTrafikskola	-0.2764	0.0840	0.001008
Dagsöndag:UtbTrafikskola	-0.0865	0.5413	0.872999
SasongSommar:Latitud	-0.0175	0.0059	0.003187
SasongVår:Latitud	-0.0066	0.0061	0.274679
SasongVinter:Latitud	-0.0106	0.0061	0.084086
Dagtisdag:Latitud	0.0014	0.0067	0.825039
Dagonsdag:Latitud	-0.0008	0.0066	0.901691
Dagtorsdag:Latitud	-0.0151	0.0068	0.027345
Dagfredag:Latitud	0.0017	0.0077	0.815659
Daglördag:Latitud	-0.0693	0.0290	0.016753
Dagsöndag:Latitud	0.0158	0.1332	0.905492

Tabell 8: Modell som minimerar AIC, ModellAIC

ModellAIC	Skattning	Std.Av	p-värde
Alder:Starttid	0.0003	0.0002	0.059142
UtbTrafikskola:Latitud	-0.0067	0.0045	0.136057
Starttid:Latitud	0.0013	0.0009	0.139689

Tabell 9: Modell som minimerar BIC, modellBIC

ModellBIC	Skattning	Std.Av	p-värde
(Intercept)	-4.6431	0.4129	< 2e-16
KonMan	0.6242	0.2577	0.015445
Alder	0.0344	0.0151	0.023354
UtbTrafikskola	1.1626	0.0129	< 2e-16
Starttid	-0.0171	0.0017	< 2e-16
Latitud	0.0921	0.0069	< 2e-16
KonMan:Alder	0.0167	0.0010	< 2e-16
KonMan:UtbTrafikskola	0.1661	0.0181	< 2e-16
Alder:Latitud	-0.0012	0.0002	6.13e-07
KonMan:Latitud	-0.0165	0.0043	0.000147

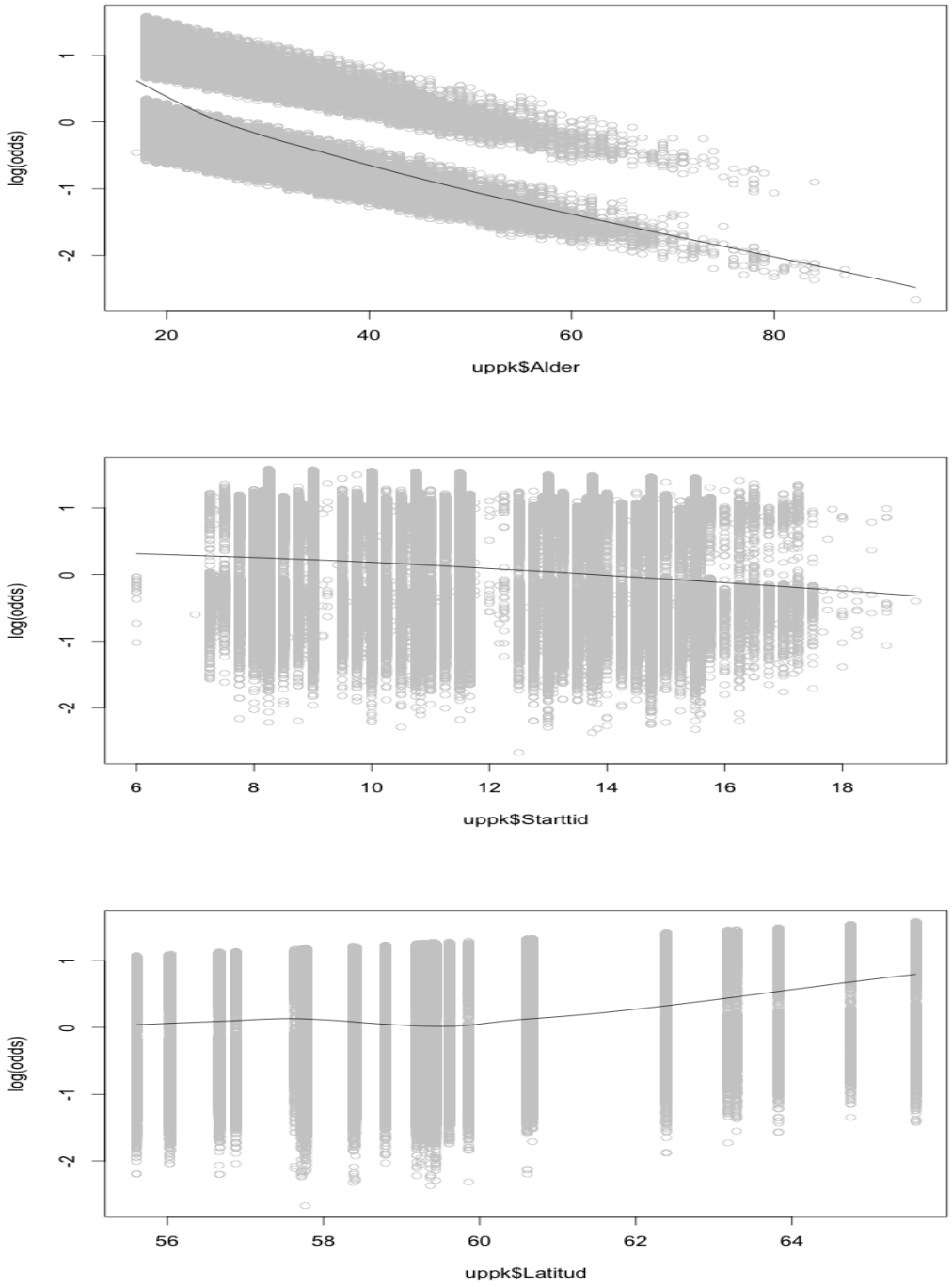
Tabell 10: Modell som är en enkel version av BIC utan samspelstermer.

ModellTest	Skattning	Std.Av	p-värde
(Intercept)	-2.5390802	0.1291916	<2e-16
KonMan	0.1367008	0.0087645	<2e-16
Alder	-0.0317584	0.0005034	<2e-16
UtbildareTrafikskola	1.2429058	0.0090992	<2e-16
Starttid	-0.0172256	0.0017866	<2e-16
Latitud	0.0517988	0.0021578	<2e-16

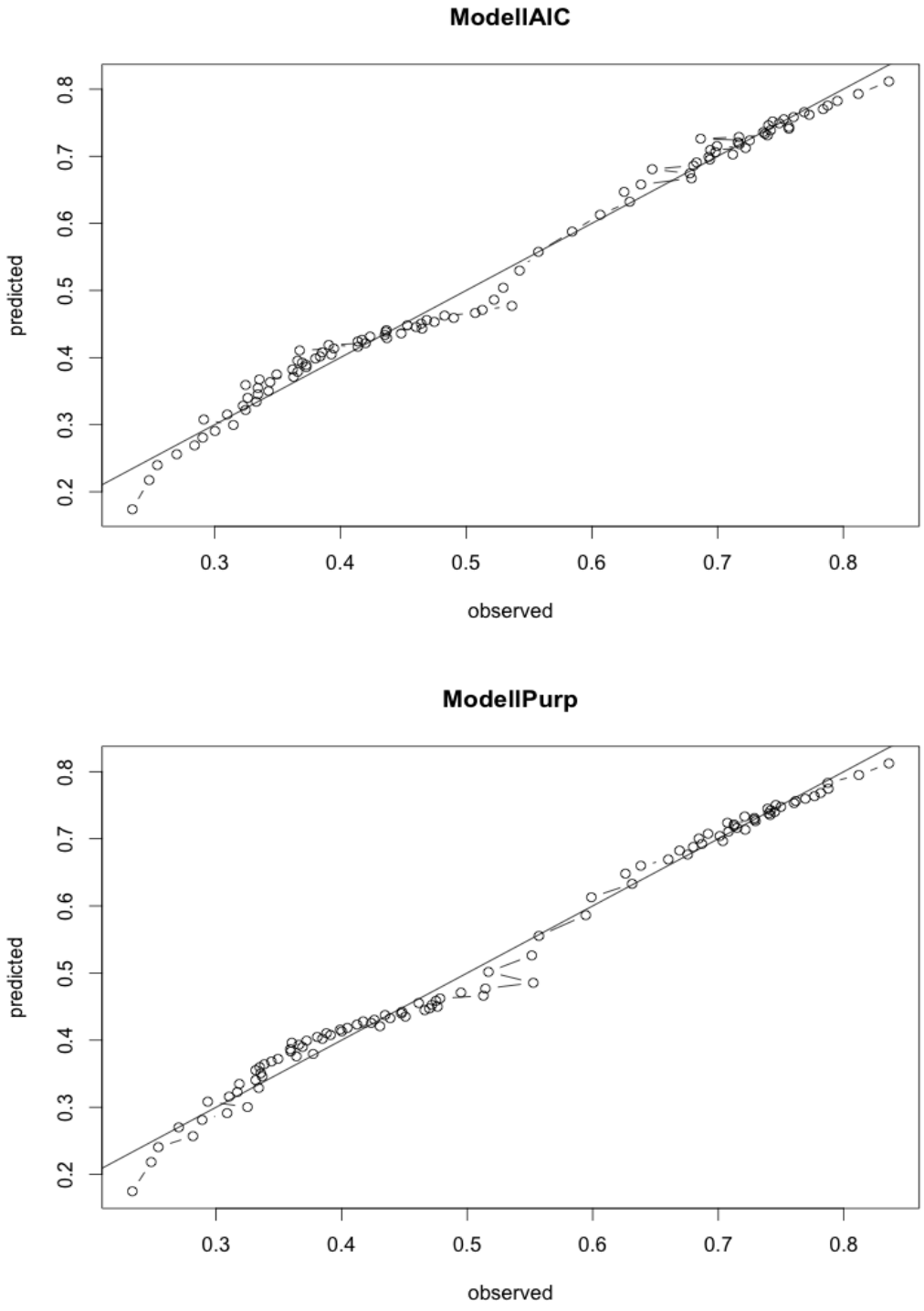
Tabell 11: Odds-ratios för modellTest.

ModellTest	Odds-Ratio	95 % Konfidens-intervall
(Intercept)	0.07893897	0.0612738, 0.1016746
KonMan	1.14648502	1.1269624, 1.1663532
Alder	0.96874056	0.9677844, 0.9696960
UtbildareTrafikskola	3.46566930	3.4044399, 3.5280618
Starttid	0.98292187	0.9794861, 0.9863697
Latitud	1.05316382	1.0487213, 1.0576296

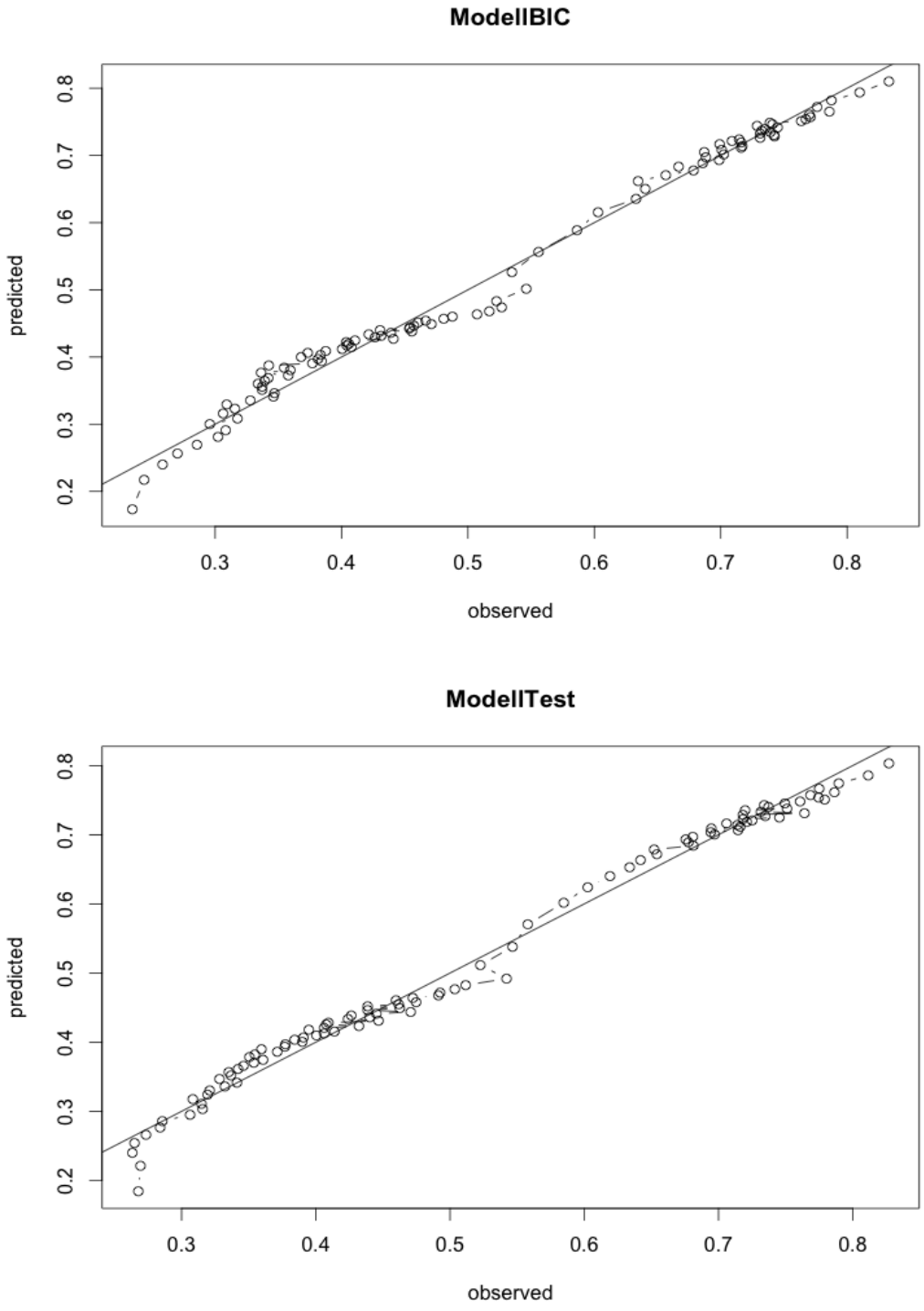
B Figurer



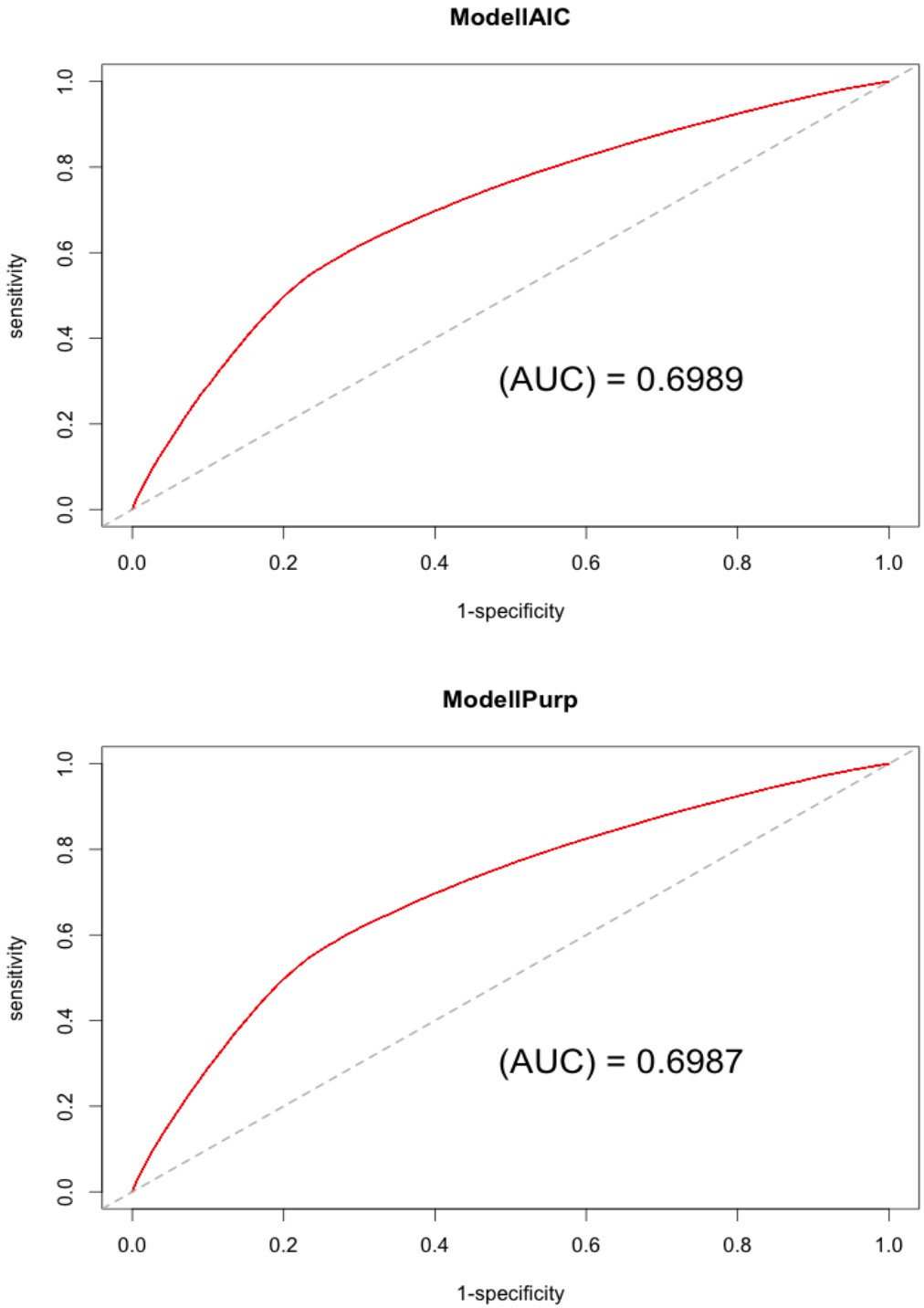
Figur 3: Lowessplottar för att undersöka linjäriteten för variabler i logiten.



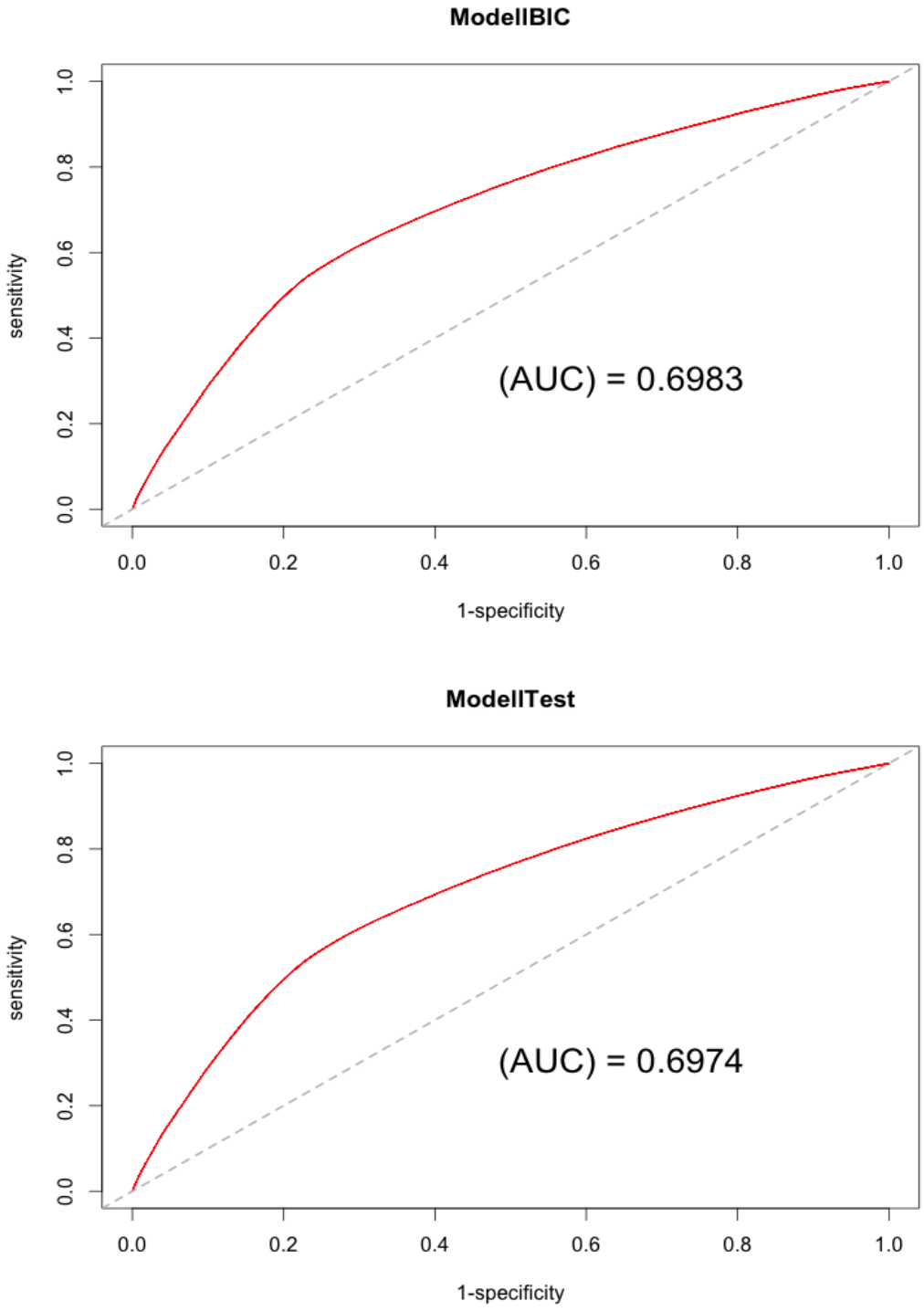
Figur 4: Observerade mot predikterade för modellAIC och modellPurp.



Figur 5: Observerade mot predikterade för modellBIC och modellTest.



Figur 6: ROC för modellAIC och modellPurp, med inkluderande AUC-värde.



Figur 7: ROC för modellBIC och modellTest, med inkluderande AUC-värde.