Stockholms universitet

Sannolikheten att klara körprovet för motorcykel - en logistisk regressions-analys

Sofie Asp

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2020:2
Matematisk statistik
Januari 2020

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2020:1**
http://www.math.su.se/matstat

Sannolikheten att klara körprovet för motorcykel - en logistisk regressionsanalys

Sofie Asp*

Januari 2020

Sammanfattning

Denna studie syftar till att förklara sannolikheten att klara körprovet för motorcykel utifrån ett antal förklarande variabler. Målsättningen är att ta fram en modell som predikterar resultatet genom logistisk regression. Modellerna tas fram med Purposeful Selection och Stepwise inkludering/exkludering för AIC, både med- och utan samspelseffekter. Modellernas goodness of fit samt predikativa förmåga undersöks med bland annat Hosmer-Lemeshow test, R-squared och ROC kurvor. Resultatet är att ingen av modellerna uppfyller villkoren för att prediktera om en individ kommer klara körprovet eller inte. Däremot bidrar studien till förståelse för vilka variabler som kan ha en signifikant inverkan på körprovsresultatet. Studien visar på att chansen för att klara uppkörningen ökar för unga män som kör upp via trafikskola i norra Sverige. För vidare studier kan tillgång till fler betydelsefulla variabler kunna generera ett mer tillfredsställande resultat.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: sea.asp@gmail.com. Handledare: Mia Deijfen.

Abstract

This study aims to explain the probability of succeeding with the driving licence test for motorcycles, using a number of explanatory variables. The goal is to present a model to predict the outcome. Analyisis are made using logistic regression. Models are constructed using purposeful selection and stepwise AIC, for each method we present models with and without interactions. The goodness of fit and the models' predictive power will be examined using Hosmer-Lemeshow tests, R-squared and ROC curves. The result is that no model holds for predictive use. Though, this study still contribute to increased knowledge for which variables may have a significant effect on the result of succeeding with the drivers test for motorcycles. We found that the chanses of succeeding increases if the driver is young, male, submitted by a driving school and doing the test in the north of Sweden. Access to additional significant variables may give a more sufficient result in future studies.

Förord

Jag vill börja med att tacka min handledare Mia Deijfen för vägledning och stöd i arbetet. Jag vill även tacka trafikverket för tillgång till datamaterialet och SMC för att hjälpa mig få kontakt med rätt person. Tack till Thomas Hjert vars examensarbete bidragit till min inspiration av ämnesval.

Innehåll

1 Inledning **5**
1.1 Bakgrund . . . 5
1.2 Syfte . . . 5

2 Metod **5**
2.1 Beskrivning av data . . . 5
2.2 Tillvägagångssätt . . . 6

3 Teori **7**
3.1 Logistisk regression . . . 7
3.2 Test av linearitet . . . 7
3.3 AIC . . . 7
3.4 Likelihood ratio test . . . 8
3.5 Generaliserande R-squared . . . 8
3.6 Hosmer-Lemeshow test . . . 8
3.7 ROC kurva . . . 9

4 Metod **9**
4.1 Modellantaganden och korrelationer . . . 9
4.2 Purposefull selection . . . 10
4.3 Stepwise AIC . . . 10

5 Modellframtagning **11**
5.1 Outliers . . . 11
5.2 Multikolinearitet . . . 11
5.3 Purposeful selection . . . 11
5.4 AIC utan samspelseffekter . . . 13
5.5 AIC med samspel . . . 13

6 Modellutvärdering **14**
6.1 VIF . . . 14
6.2 Hosmer Lemeshow test . . . 14
6.3 R-squared . . . 15
6.4 Receiver Operating Characteristics . . . 15
6.5 Tolkning av koefficienterna . . . 15

7 Diskussion **15**

8 Slutsats **16**

9 Källförteckning **17**

10 Bilaga **18**
A Grafer . . . 18
B Tabeller . . . 21

1 Inledning

1.1 Bakgrund

I Sverige utförs uppkörningar för motorcykel av Trafikverket. Körprovet är uppdelat i flera moment: en säkerhetskontroll, krypkörning, högfartsbana, inbromsning i hög fart och därefter körning i trafik om samtliga moment är godkända. Vid misslyckande på någon av manöverproven utförs vanligtvis inte körningen i trafik. Att klara en uppkörning för motorcykel kan således verka svårare än att klara uppkörning för bil. Enligt SMC (u.å.) är det svårare för kvinnor att klara manöverbanorna på grund av att kvinnor generellt sett är kortare och väger mindre än män. Den 1 maj 2015 genomförde Transportstyrelsen en ändring av körproven för motorcyklar som ska göra det lättare för kvinnor att klara uppkörningen (Transportstyrelsen, 2015). Det finns tre olika klasser på motorcyklar: A1 (lätt) som man får köra upp för från 16 års ålder, A2 (tung med begränsning) som man får köra upp för från 18 års ålder och A (tung utan begränsning) som man numera inte får köra upp för förrän man fyllt 24 år. Om man har haft A2 i minst 2 år får man köra upp för A vid 20 års ålder.

Privat övningskörning kräver en handledare som innehavt behörigheten i minst 5 år och måste åka med på egen eller samma motorcykel (Mc-jakten, u.å.). Det är inte alla som har tillgång till en privat handledare och många som vill ta motorcykelkörtkort är tvungna att vända sig till kostsamma trafikskolor.En vedertagen uppfattning bland gemene man är att det är lättare att klara uppkörningar i mindre orter än i storstäder, och även mer norrut i landet. Denna studie kommer bland annat undersöka om dessa faktorer har någon signifikant inverkan på om en individ klarar uppkörningen eller inte.

1.2 Syfte

Syftet med denna studie är att utifrån data från Trafikverket studera vilka förklarande variabler som har en signifikant inverkan på om en individ kommer klara provet eller inte. Vi undersöker även om de modeller som tas fram kan användas för att prediktera resultatet.

2 Metod

2.1 Beskrivning av data

Vi har erhållit data på 55 075 st uppkörningar för motorcyklar i hela landet från Trafikverket. Uppkörningarna sträcker sig under tidsperioden 2016-2019, vilket innebär att samtliga uppkörningar som behandlas i rapporten är genomförda med samma provregler. Vi har även lagt till data som beskriver uppkörningsortens storlek, dels i antal invånare i tätorten och dels om det är en stor stad (>50 000 invånare) eller inte.

Tabell 1: Samtliga variabler

Variabel	Beskrivning av variabeln	Typ	Nivåer
Ålder	Ålder vid uppkörning	Kontinuerlig	16-75
Kön	Man/kvinna	Kategorisk	0=Kvinna, 1=Man
Provtyp	Lätt-, mellanklassad- eller tung motorckel	Kategorisk	A1 = lätt, A2 = mellan, A = tung
Datum	Datum för provtillfället	Kontinuerlig	dag/mån/år
Starttid	Tid provet började	Kontinuerlig	6.00-21.40
Förmiddag	Förmiddags-/eftermiddagsprov	Kategorisk	Förmiddag/eftermiddag
Veckodag	Dag i veckan	Kategorisk	Måndag-Söndag

Variabel	Beskrivning av variabeln	Typ	Nivåer
Månad	Månad på året	Kategorisk	Januari-December
Säsong	Årstid	Kategorisk	Vår/sommar/höst/vinter
Utbildare	Om man kört upp privat eller genom trafikskola	Kategorisk	0=Privatist, 1= Trafikskola
Resultat	Om man klarat provet eller inte	Kategorisk	0=Underkänd, 1=Godkänd
Kontor	På vilket kontor uppkörningen sker	Kategorisk	
ProvplatsNamn	Vad provplatsen heter	Kategorisk	
ProvplatsPostnr	Vilket postnummer provplatsen har	Kategorisk	
ProvplatsOrt	Vilken ort uppkörningen skett i	Kategorisk	
ProvplatsLat	Latitud, där ju mer norrut desto högre värden	Kontinuerlig	55.45-65.84
ProvplatsLong	Longitud hur mycket västerut/österut	Kontinuerlig	11.91-23.29
ProvplatsKoord	Koordinaterna för provplatsen	Kontinuerlig	latitud:longitud
Befolk	Antal invånare i tätorten år 2018	Kontinuerlig	47-317 245
Stad	Om det klassas som en stor eller liten stad	Kategorisk	0=liten, 1=stor

De variabler som beskrivits som kontinuerliga skulle snarare kunna ses som kategoriska, men på väldigt många nivåer.

Korrelationer i datan undersöks vidare i avsnittet "Multikolinearitet" (se avsnitt 5.2). I modellframtagningen undersöks även linearitetens av de kontinuerliga variablerna, vilket resulterar i omkonstruktion av variabeln som beskriver Latitud. Den slutgiltiga tabellen efter exklusion av variabler som inte ingår i någon av slutmodellerna blir:

Tabell 2: Variabler som ingår i slutmodellerna

Variabel	Beskrivning av variabeln	Typ	Nivåer
Ålder	Ålder vid uppkörning	Kontinuerlig	16-75
Kön	Man/kvinna	Kategorisk	0=Kvinna, 1=Man
Provtyp	Lätt-, mellanklassad- eller tung motorckel	Kategorisk	A1 = lätt, A2 = mellan, A = tung
Förmiddag	Förmiddags-/eftermiddagsprov	Kategorisk	Förmiddag/eftermiddag
Veckodag	Dag i veckan	Kategorisk	Måndag-Söndag
Säsong	Årstid	Kategorisk	Vår/sommar/höst
Utbildare	Om man kört upp privat eller genom trafikskola	Kategorisk	0=Privatist, 1=Trafikskola
Resultat	Om man klarat provet eller inte	Kategorisk	0=Underkänd, 0=Godkänd
Lat_Kat	Latitud, indelat efter 20e precentiler	Kategorisk	55.44-56.86, 56.95-57.79, 57.81-59.17, 59.23-5
Befolk	Antal invånare i tätorten år 2018	Kontinuerlig	47-317 245
Stad	Om det klassas som en stor eller liten stad	Kategorisk	0=Liten, 1=Stor

2.2 Tillvägagångssätt

Den statistiska undersökningen tillämpar logistisk regressionsanalys i programmet R med hjälp av den generaliserade linjära modellen (glm). Glm använder sig av maximum likelihood, till skillnad från linjär regression som använder sig av minsta kvadrat metoden för att skatta parametrar. Först bestäms vilka kovariater som ska tas med i modellen och även deras parameterskattningar genom två olika metoder; *purposeful selection* och *stepwise selection AIC* (se avsnitt 4). Modellens *goodness of fit* bedöms genom att undersöka likelihood ratio tester och Hosmer-Lemeshow test. Därefter undersöks modellens predikativa förmåga med hjälp av Recievers Operating Characteristics Curve (ROC) och tre olika pseudo R^2. Vi studerar även β koefficienterna och tolkar oddskvoterna för en av modellerna.

3 Teori

I detta avsnitt görs en kortfattat genomgång av logistisk regeression samt presenterar de tester som tillämpas i analysen.

3.1 Logistisk regression

Logistisk regression är enligt Agresti (2013) den viktigaste modellen för kategorisk responsdata. I en multipel logistisk regressionsmodell förklaras responsvariabeln Y som, kan anta binära värden med flera förklaringsvariabler som benämns med $\mathbf{x} = (x_1, ..., x_p)$

Den logistiska regressionsmodellen är

$$\pi(\mathbf{x}) = exp(\alpha + \beta x)/1 + exp(\alpha + \beta \mathbf{x}),$$

där $\pi(\mathbf{x}) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x)$ och $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$.

Oddsen definieras som

$$\Omega = \pi(\mathbf{x})/(1 - \pi(\mathbf{x})) = exp(\alpha + \beta \mathbf{x}).$$

Logaritmering av oddset innebär att man tar logaritmen av sannolikheten delat på 1-sannolikheten och detta kallas även för logiten av sannolikheten. Den kan anta alla reella tal medan sannolikheten $\pi(\mathbf{x})$ endast kan anta värden mellan 0 och 1. Logaritmering av oddsen ger det linjära sambandet

$$log(\pi(\mathbf{x})/(1 - \pi(\mathbf{x})) = logit[\pi(\mathbf{x})] = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p.$$

Regressionskoefficienterna β berättar hur logiten för händelsen ändras när x ökar med 1 enhet. Till exempel om $\beta = 0.5$ så innebär det att när x ökar med 1 enhet så ökar logiten för att händelsen ska inträffa med 0.5. Detta kan omvandlas till ett mer lättolkat format genom att beräkna oddskvoten: $exp(0.5) = 1.65$. Nu kan vi tolka detta som att när x ökar med en enhet så ökar oddset för händelsen med 65 %. Om oddskvoten är 1 innebär det att det inte finns något samband mellan responsen och den förklarande variabeln x.

3.2 Test av linearitet

De kontinuerliga variablerna i modellen behöver testas för linearitet mot logiten. Detta utförs genom att dela upp den kontinuerliga variabeln som ska testas i fem delar efter dess 20:e precentiler och därefter byter ut dessa som "kategoriska variabler" mot den kontinuerliga ursprungsvariabeln. Parameterskattningarna för varje del bör ligga nära respektive medelvärde. Om variabeln inte är linjär i logiten kan den antingen transformeras eller bytas ut som kategorisk variabel (Hosmer & Lemeshow, 2000).

3.3 AIC

Akaike Information Criterion är ett mått på hur pass mycket modellen missar. Den hanterar både under- och överanpassning av modellen.

$$AIC = -2l(\hat{\theta}_{ML}) + 2p$$

AIC använder sig av maximum log-likelihood, där den belönar *goodness of it* och straffar för fler parametrar (Applied Statistical Inference, 2014, s. 224). Ju lägre AIC desto bättre modellanpassning. Sålunda väljs den modell med lägst värde på AIC.

3.4 Likelihood ratio test

För att mäta *goodness of fit* används likelihood-ratio tester. Då jämförs en enklare modeller med en mer komplex modell. Den komplexa modellen är den ursprungliga modellen M_0 och den enklare modellen är den alternativa modellen M_A. Om den mer komplexa modellen ger en bättre anpassning så väljs den ursprungliga modellen. I *likelihood ratio* tester beräknas loglikelihooden för respektive modell och jämförs om skillnaden är signifikant. I modellframtagningen används funktionen *lrtest* som även berättar om skillnaden mellan modellerna är signifikant, och om den alternativa modellen bör väljas framför den ursprungliga.

$$LR = -2log(L_0(\theta)/L_A(\theta))$$

3.5 Generaliserande R-squared

I linjär regression anger R^2 statistikan hur stor del av variansen som förklaras av de oberoende variablerna. Förklaringsgraden ges av

$$R^2 = 1 - \Sigma_i(y_i - \hat{y}_i)^2/\Sigma_i(y_i - \overline{y})^2,$$

där y_i är det observerade värdet, $\hat{y}_i$ vårskattning och $\overline{y}$ medelvärdet.

I logistisk regression finns det inte någon exakt motsvarighet, men det finns en del R^2 värden som kan vara värdefulla. En av dessa är McFadden's R^2 som definieras som

$$R_{McF} = 1 - ln(L_M)/ln(L_0),$$

där $ln(L_M)$ är log likelihooden för vår anpassade modell och $ln(L_0)$ är log likelihooden för nollmodellen. Detta mäts med ett värde mellan 0 och 1, där ju närmare 0 desto sämre prediktionskraft.

Ett annat vanligt mått på R^2 är Cox och Snells (1989):

$$R_{CS} = 1 - ((L_0)/(L_M))^{(2/n)}.$$

Om vi dividerar R_{CS} med dess övre gräns erhåller vi ett tredje mått på R^2 som kallas för Nagelkerke (1991):

$$R_{CS} = (1 - ((L_0)/(L_M))^{(2/n)})/(1 - {L_0}^{2/n}).$$

För n antal observationer tar vi den n:te roten ur likelihooden för nollmodellen dividerat på likelihooden för vår anpassade modell. Detta ska ge oss ett skattat värde på likelihooden för varje Y. Observera att Cox och Snells modell har ett maxvärde lägre än 1.

I R använder vi oss av funktionen *PseudoR2* och kollar på värdet för "McFadden", "Cox & Snell" samt "Nagelkerkes" R^2.

3.6 Hosmer-Lemeshow test

Hosmer-Lemeshow test använder sig av Pearson χ^2 statistikan för att jämföra observerade och skattade värden med varandra. Pearson χ^2 *goodness of fit test* är en användbar metod om det finns många observationer för varje x-variabel.

Testet går ut på att dela upp urvalen utifrån deras skattade sannolikheter. För varje koefficientskattning beräknas sannolikheten att $Y = 1$ för varje kovariat

$$\hat{\pi} = exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + ... + \hat{\beta}_p X_p)/1 + exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + ... + \hat{\beta}_p X_p).$$

Vi delar in observationerna i $g = 10$ grupper. Den första gruppen består av de observationer som har de lägsta 10 % skattade sannolikheterna. Nästa grupp består av de näst lägsta 10 % skattade sannolikheterna osv.

Vidare beräknas hur många $Y = 1$ observationer vi kan förvänta oss, genom att ta medelvärdet av de skattade sannolikheterna multiplicerat med antalet observationer i gruppen. Detsamma görs för $Y = 0$ och därefter beräknas en Pearson *goodness of fit* statistika

$$\Sigma_{k=0}^{1}\Sigma_{l=1}^{g}(o_{kl} - e_{kl})^2/e_{kl}.$$

Här anger o_{0l} antal observationer $Y = 0$ i den l:the gruppen, och o_{1l} anger antal observerade $Y = 1$ i den l:the gruppen. e_{0l} och e_{1l} anger det förväntade antalet nollor.

För att genomföra testet använder vi funktionen *hoslem.test* i R som genererar ett värde på χ^2 med $g - 1$ frihetsgrader. Ett litet p-värde indikerar att modellen ger en dålig anpassning till datan. χ^2-värdet blir större vid stort antal observationer (Hosmer & Lemeshow, 2000, s. 147). Eftersom vi har ett stort dataset kan vi förvänta oss signifikans. Testresultatet kommer inte enskilt kunna ge oss den information som behövs men kommer användas som ett komplement till övriga tester och jämförelse av χ^2 mellan modellerna.

3.7 ROC kurva

Recieving operating charecteristic (ROC) baseras på proportionerna av positiva datapunkter TPR (True Positive Rate) och proportionen av negativa datapunkter TNR (True Negative Rate). TRP är andelen fall där vi predikterat ett positivt utfall och det verkliga utfallet är positivt. På liknande sätt motsvarar TNR andelen fall där vi predikterat ett negativt utfall och det verkligen utfallet är negativt. Vi har även något som kallas för FPR (False Positive Rate) när vi predikterat ett positivt utfall, men det verkliga utfallet är negativt.

Dessa mått visas grafiskt i en ROC kurva, där vi plottar sensitiviteten mot 1-specificiteten och binder ihop punkterna (0,0) och (1,1).

$$sensitivitet = TPR = TP/(TP + FN)$$
$$1 - specificitet = FPR = FP/(TN + FP)$$

Prediktionsmåttet anges av arean under ROC kurvan och benämns som AUC. AUC:n ger ett mått som går mellan 0 och 1, där desto närmare 1 ju bättre anses modellens prektionskraft vara (Hosmer & Lemeshow, 2000, s. 160). Det vill säga; ju större arean är, desto starkare är modellens prediktionskraft. Arean är ett mått på sannolikheten att vår modell predikterar ett positivt resultat. Det vill säga att om $AUC = 1$ så får vi 100 % rätt när vi försöker prediktera om resultatet är 1 eller 0.

Om $AUC = 0.5$ är prediktionskraften inte bättre än slumpmässiga gissningar (Agresti, 2013). Om $0.5 < AUC < 0.7$ är det inte mycket bättre än gissningar, dvs dåligt. Om $0.7 < AUC < 0.8$ så har vi ett acceptabelt mått. Om $0.8 < AUC < 0.9$ anses det som ett bra mått och vid $AUC \geq 0.9$ så har vi ett väldigt bra mått. Det är dock ytterst ovanligt att AUC:n är större än 0.9 (Hosmer & Lemeshow, 2000, s. 162).

I en ROC vill vi ha en kurva som är så nära det övre vänstra hörnet i grafen, och så långt bort ifrån den 45 gradiga baslinjen, där FPR=TPR, då det indikerar på låg prediktionskraft.

4 Metod

Vi kommer ta fram modeller med hjälp av två olika metoder; purposeful selection och stepwise AIC. Tillvägagångssättet presenteras i detta avsnitt.

4.1 Modellantaganden och korrelationer

Binär logistisk regression förutsätter att den beroende variabeln är binär. Det krävs även att observationerna är oberoende av varandra samt att de förklarande variablerna får inte starkt korrelera. I de fall korrelationen ter sig starkt väljs det att utesluta ett antal variabler och konstruera om andra.

4.2 Purposefull selection

Detta är en metod som beskrivs av Hosmer och Lemeshow (2000).

Steg 1: Enskilt undersöka respektive förklarande variabel med hjälp av envariabels logistisk regression. Om den förklarande variabeln genererar ett icke-signifikant p-värde > 0.25 exkluderas den ur modellen.

Steg 2: Vi studerar en multivariat modell med samtliga variabler som inte förkastats i föregående steg. I detta steg vill vi att variablerna ska p-värde på signifikansnivå 0.05, annars exkluderas dem. Därefter jämförs modellen med den tidigare genom ett partiellt likelihood ratio test.

Steg 3: Nu jämförs koefficientskattningarna i båda modellerna och observerar om koefficienten för någon variabel i den mindre modellen har en skillnad på mer än 20 % jämfört med den större. Om så är fallet har de variabler som tagits bort en stor inverkan på andra variablers skattningar, och de bör tas tillbaka i modellen. Vi går igenom steg 2 och 3 tills alla relevanta variabler ingår och alla variabler som har exkluderats är statistiskt icke-signifikanta. Hosmer och Lemeshow (2000) använder ett mått för förändringen av koefficienterna som kallas för "delta-beta-hat-percent":

$$\Delta\hat{\beta} = 100(\hat{\theta}_1 - \hat{\beta}_1)/\hat{\beta}_1$$

där $\hat{\theta}_1$ är koefficienten för den mindre modellen och $\hat{\beta}_1$ är koefficienten för den större modellen.

Steg 4: I detta steg vill vi se om det finns variabler som bidrar till samvariation och som i sig inte är signifikanta. Vi lägger till alla variabler som exkluderats i steg 1 och kontrollerar signifikansen.

Efter detta steg är vi kvar med en modell som kallar *preliminära huvudeffektsmodellen*.

Steg 5: Här undersöks om nivåerna för de kategoriska variablerna är rimliga och att de kontinuerliga variablerna är linjära i logiten.

Modellen efter detta steg kallas *huvudeffektsmodellen*.

Steg 6: Vidare undersöks om eventuella samspel mellan variabler bör ingå i modellen. För att göra detta läggs ett samspel till i taget och vi kollar på den statistiska signifikansen av samspelet. Om samspelet är signifikant så läggs den till i modellen. Vi återgår till steg 2 men kollar enbart på samspelseffekterna. Vi kommer alltså inte ta bort någon av huvudeffekterna denna gång.

Modellen kallas nu *preliminär slutmodell*".

Steg 7: Innan vi når vår slutgiltiga modell testar vi den för flera olika *goodness of fit tester*.

4.3 Stepwise AIC

Genom stegvis inkludering/exkludering av kovariat väljs den modell som ger lägst AIC. Proceduren baseras på en kombination av framåtinkludering och bakåteliminering. Framåtinkludering innebär att att vi börjar med en modell utan variabler och lägger stegvis till kovariat en och en, och utvärderar om de ska tas med i modellen eller inte. I bakåteliminering börjar vi med en full modell och tar stegvis bort ett kovariat och utvärderar om den ska behållas eller elimineras i modellen.

Vi tar hjälp av R's funktion *step* där vi tar fram en modell med hjälp av både inkludering och exkludering för att hitta vilka kovariat som ger lägst AIC. Vi tar först fram en modell utan samspel, och därefter lägger till möjliga samspel och utför stegvis inkludering/exkludering igen för att få fram en modell med samspel.

5 Modellframtagning

5.1 Outliers

Det finns endast en observation från december och en observation från mars. Motorcykelsäsongen sträcker sig vanligtvis från april till mitten av november, därav exkluderas dessa två observationer. Variabeln 'Säsong' kommer därmed endast bestå av tre nivåer: vår, sommar och höst.

5.2 Multikolinearitet

Vi börjar med att studera datamaterialet och finna vilka kovariat som har en signifikant inverkan på resultatet, samt hur eventuella samvariationer ska hanteras. Den givna datan innehåller bland annat information om kontor, namn på provplatsen, ort, postnummer, latitud, longitud samt koordinater, där alla beskriver "var" uppkörningen har skett och kan sålunda antas korrelera. Vi kan enkelt utesluta variabeln 'koordinater' i modellen eftersom den består av variablerna 'longitud' och 'latitud'. Postnummer ger liknande information som provplatsens ort, namn och kontor. Alla dessa variabler ger oss information om var i landet uppkörningen har skett. Det är överflödigt att inkludera variabler som ger samma eller väldigt liknande information. Vi väljer att ta med variablerna 'longitud' och 'latitud' som förklarar var i landet uppkörningen skett.

Om vi ska använda både variabeln 'ProvplatsLat' och 'ProvplatsLong' behöver vi se så dessa inte är starkt korrellerade. Vi får fram att korrelationskoefficienten är 0.7144, vilket innebär att det finns en positiv korrelation. Sålunda väljer vi att inkludera den variabel som enskilt bäst förklarar vår respons och exkludera den andra.

Det råder även tydligt samband mellan 'Tid' och 'Förmiddag', samt mellan 'Månad' och 'Säsong'. 'Förmiddag' förklarar resultatet något bättre än 'Tid' och sålunda väljer vi att inkludera den variabeln i analysen. En motorcykelsäsong i Sverige varar mellan april och mitten av november. Det är därmed rimligare att kolla på variabeln säsong än månad.

5.3 Purposeful selection

Steg 1: Vi genomför enkel logistisk regression för samtliga variabler. De som uppfyller kravet på $p < 0.25$ behålls kvar till steg 2 och presenteras i tabell 12 i bilagan.

Steg 2: Vi skapar en multivariat modell med samtliga signifikanta kovariat från steg 1 och undersöker om signifikansen håller med krav på $p \leq 0.05$.

Observera att för de kategoriska variablerna ingår den första nivån i interceptet. Vi ser att inte samtliga nivåer är signifikanta i variabeln 'Veckodag' men eliminering av dessa skulle ge oss icke-väntevärdesriktiga skattningar. Om man ska exkludera kategoriska variabler så behöver samtliga nivåer exkluderas (Hosmer & Lemeshow, 2000, s. 38). Det framgår att variabeln 'Befolk' inte är signifikant i den multivariata modellen och således exkluderas den.

Vi jämför därefter de två modellerna med ett likelihood ratio test och får att modell två är signifikant på nivån 0.001 vilket innebär att vi kan förkasta nollhypotesen om att modell 1 skulle vara bättre. Studerar vi på AIC så ser vi att den skiljer sig från 65490 i modell 1 till 65482 i modell 2. Vi väljer att gå vidare med modell 2.

Steg 3: Vidare studeras koefficienterna och vi finner inga drastiska förändringar mellan den ursprungliga och den reducerade modellen. Vi går vidare till steg 4 där vi en i taget lägger till varje variabel som inte valdes med från steg 1. Vi undersöker signifikansen och om de fortfarande inte är signifikanta så utesluts dem. Om de nu visar sig vara signifikanta kan det råda oväntade samvariationer som behövs åtgärdas. Vi finner att de fortfarande är icke-signifikanta och sålunda går processen vidare till Steg 5.

Steg 5: I detta steg ska vi undersöka om de kontinuerliga variablerna är linjära i logiten och om kategoriseringen i de kategoriska variablerna är bra konstruerade. Resultatet presenteras i graf 1 och 2 i bilagan. Vi genomför även ett box-Tidewell test för att stärka resultatet av plottarna.

Tabell 3: Box-Tidewell test av linearitet

Variabel	MLE of lambda	Score Statistic (z)	Pr(>|z|)
Ålder	-0.0572	1.5627	0.1181
ProvplatsLat	-2.6377	-2.5533	0.0107 *

I boxTidwell-testet söker vi icke-signifikanta resultat, som ska indikera på att variablerna är lineära logiten. Ålder ger ett icke-signifikant resultat, vilket innebär att den kan antas vara lineär i logiten. Detta stämmer överens med vad vi ser i graf 1 (se bilaga). Däremot är 'ProvplatsLat' signifikant på nivån 0.5, och grafen visar inte på linearitet. Vi kommer sålunda kategorisera variabeln 'ProvplatsLat' i vår modell enligt tabellen nedan.

Tabell 4: Kategorisering av latitud

Variabelnamn	Beskrivning	Nivåer
Lat_Kat20	Södra Sverige (ca Malmö-Halmstad)	55.4454-56.8602
Lat_Kat40	Södra Sverige (ca Växjö-Göteborg)	56.9544-57.7997
Lat_Kat60	Mellersta Sverige (ca Jönköping-Södertälje)	57.8127-59.1729
Lat_Kat80	Mellersta Sverige (ca Stockholmsområdet)	59.2336-59.6552
Lat_Kat100	Norrland (ca Uppsala-Luleå)	59.9234-65.8368

Vi undersöker den kategoriserade variabeln med logistisk regression och finner att 'Lat_Kat' är en signifikant variabel i modellen. Vi ser även att AIC:n nu är 65401, till skillnad från innan 65482. Detta indikerar att kategoriseringen av latituden ger en bättre anpassad modell.

Tabell 5: Purposeful Selection utan samspel

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.6063	0.0638	9.5091	0.0000
Kön	0.3520	0.0272	12.9391	0.0000
Ålder	-0.0266	0.0011	-24.5256	0.0000
ProvtypKörprov A1	-0.9119	0.0370	-24.6333	0.0000
ProvtypKörprov A2	-0.2826	0.0313	-9.0140	0.0000
Förmiddag	0.0754	0.0201	3.7415	0.0002
VeckodagLördag	0.0184	0.0690	0.2659	0.7903
VeckodagMåndag	0.1340	0.0406	3.3032	0.0010
VeckodagOnsdag	0.1021	0.0377	2.7081	0.0068
VeckodagSöndag	0.0182	0.2260	0.0806	0.9358
VeckodagTisdag	-0.0012	0.0366	-0.0319	0.9745
VeckodagTorsdag	0.1113	0.0371	2.9997	0.0027
SäsongSommar	0.0394	0.0225	1.7497	0.0802
SäsongVår	0.1132	0.0291	3.8875	0.0001
Utbildare	0.9201	0.0199	46.1230	0.0000
Lat_Kat40	-0.0475	0.0297	-1.5990	0.1098
Lat_Kat60	0.1534	0.0301	5.0872	0.0000
Lat_Kat80	0.2277	0.0299	7.6096	0.0000
Lat_Kat100	0.5961	0.0331	18.0163	0.0000
Stad	0.0000	0.0000	3.2058	0.0013

Steg 6: Nu undersöker vi samspel, genom att en och en lägga till möjliga samspel i modellen och studera dess signifikans.

Efter vi funnit signifikanta samspel återgår vi till steg 2 och inkluderar samtliga samspel som uppvisat signifikans. Vi exkluderar nu de variabler som inte längre uppfyller ett p-värde på signifikansnivå 0.05 och utför ett *likelihood ratio test*.

Modellen efter exkludering av icke-signifikanta samspel presenteras i bilaga (se tabell 13).

Vi genomför ett *likelihood ratio test* som resulterar i att nollhypotesen om att den tidigare modellen utan samspel är bäst kan förkastas. Vi går vidare med modellen med samspelseffekter (tabell 13) som blir vår *preliminära slutmodell*. Vi ser även här att AIC:n har förbättrats till 64968. För att nå en *slutmodell* genomförs olika *goodness of fit* tester i avsnitt 6.

5.4 AIC utan samspelseffekter

Vi använder oss av R:s funktion *step* för att göra stegvis framåt- och bakåtselektion för att ta fram den modell som ger lägst AIC. Vi tar fram en modell både med och utan samspelseffekter.

Tabell 6: Stegvis AIC utan samspel

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.7161	0.0679	10.5418	0.0000
Kön	0.3525	0.0272	12.9532	0.0000
Ålder	-0.0266	0.0011	-24.4646	0.0000
ProvtypKörprov A1	-0.9142	0.0370	-24.6886	0.0000
ProvtypKörprov A2	-0.2839	0.0314	-9.0554	0.0000
Förmiddag	0.0633	0.0203	3.1196	0.0018
VeckodagLördag	0.0031	0.0691	0.0443	0.9647
VeckodagMåndag	0.1265	0.0406	3.1145	0.0018
VeckodagOnsdag	0.0871	0.0378	2.3013	0.0214
VeckodagSöndag	0.0128	0.2258	0.0565	0.9549
VeckodagTisdag	-0.0171	0.0368	-0.4644	0.6424
VeckodagTorsdag	0.0938	0.0373	2.5147	0.0119
SäsongSommar	0.0376	0.0225	1.6714	0.0946
SäsongVår	0.1141	0.0291	3.9151	0.0001
Utbildare	0.9252	0.0200	46.2804	0.0000
Lat_Kat40	-0.1229	0.0338	-3.6332	0.0003
Lat_Kat60	0.0851	0.0335	2.5382	0.0111
Lat_Kat80	0.1823	0.0315	5.7866	0.0000
Lat_Kat100	0.5346	0.0356	14.9995	0.0000
Befolk	0.0000	0.0000	-4.7315	0.0000
Stad	0.0000	0.0000	3.0622	0.0022

Modellen Stepwise AIC utan samspel (se nedan tabell 6) skiljer sig från den första som togs fram med purposeful selection på sådant sätt att den inkluderar variabeln 'Befolk'. Denna modell ger som förväntat något bättre AIC på 65381 jämfört med modellen i purposeful selection med AIC=65401.

5.5 AIC med samspel

Vi utgår nu från modellen AIC utan samspel och lägger till alla möjliga samspelseffekter. Vi kör funktionen i R och landar i den slutmodell som presenteras i tabell 14 i bilagan.

Det är möjligt att med större kunskap skulle ett antal samspel kunna elimineras, på grund av att de inte uppfyller villkoret om att vara realistiska. Vi kan enbart ha aningar om att vissa samspel inte är av stor betydelse, men eftersom dessa har en tydlig inverkan på resultatet väljer vi att i detta skede behålla dom i vår slutmodell.

6 Modellutvärdering

6.1 VIF

Vi börjar med att kolla så det inte råder oanade korrelationer i modellerna. Vi beräknar VIF-faktorn för respektive modell utan samspelsfaktorer, eftersom samspel bidrar naturligt till samvariation.

Tabell 7: VIF för purposeful selection utan samspel

Variabel	Df	VIF
Kön	1	1.0177
Ålder	1	1.3130
Provtyp	2	1.1373
Förmiddag	1	1.0300
Veckodag	6	1.0078
Säsong	2	1.0103
Utbildare	1	1.0546
Lat_Kat	4	1.0235
Stad	1	1.0119

Tabell 8: VIF för respektive modell utan samspel

Variabel	Df	VIF
Kön	1	1.0177
Ålder	1	1.3132
Provtyp	2	1.1374
Förmiddag	1	1.0379
Veckodag	6	1.0089
Säsong	2	1.0105
Utbildare	1	1.0566
Lat_Kat	4	1.0641
Befolk	1	1.2024
Stad	1	1.0126

I tabellerna ovan ser vi att ingen variabel har en VIF-faktor > 5, vilket innebär att det inte råder någon stark multikolinearitet i modellen.

6.2 Hosmer Lemeshow test

Om p-värdet är mindre än 0.05 så indikerar det att modellen är dålig. Modellen framtagen av purposeful selection med samspelseffekter ger ett p-värde på 0.171, vilket innebär att modellen inte nödvändigtvis ger en dålig anpassning till resultatet. Vi kan dock heller inte säga att den är bra. Alla de andra modellerna har ett p-värde$<$0.05 vilket skulle indikera på dålig anpassning, men i vårt fall behöver vi ha i åtanke att detta test ofta genererar ett litet p-värde vid stora datamängder.

Tabell 9: Hosmer-Lemeshow goodness of fit

Modell	Chi-två	df	p-värde
PS utan samspel	38.30	8	6.64e-06
PS samspel	11.58	8	0.171
Stepwise utan samspel	44.07	8	5.52e-07
Stepwise samspel	17.05	8	0.0296

6.3 R-squared

McFadden R^2 säger hur stark predikativ förmåga modellen har. Ju närmare noll, desto sämre predikativa förmåga och ju närmare 1, desto bättre. Vi får värden som är nära noll, vilket indikerar att modellerna har dålig predikativ förmåga.

Tabell 10: Anpassningsmått

Modell	AIC	McFadden	Cox&Snell	Nagelkerke
PS utan samspel	65401	0.0516	0.0626	0.0876
PS samspel	64968	0.0592	0.0714	0.1000
Stepwise utan samspel	65381	0.0519	0.0629	0.0882
Stepwise samspel	64805	0.0627	0.0755	0.1058

6.4 Receiver Operating Characteristics

Vi tar fram och plottar ROC kurvan för de fyra modellerna (se graf 3-6 i bilagan). Vi beräknar AUC:n som representerar ett mått för modellens prediktionskraft och får följande:

Tabell 11: AUC

Modell	AUC
PS utan samspel	0.6551
PS samspel	0.6648
Stepwise utan samspel	0.6556
Stepwise samspel	0.6695

Det går nu att konstatera att modellen framtagen med stegvis AIC har starkast prediktionskraft av samtliga framtagna modeller. Det är även den modell som har flest förklarande variabler. Samtliga modeller har en AUC < 0.7, vilket innebär att ingen av modellerna är tillräckligt bra för att användas i prediktionssyfte.

6.5 Tolkning av koefficienterna

Koefficientskattningarna tolkas som lättast med oddskvot. I bilagan presenteras oddskvoten för modellen Stepwise AIC i tabell 15. Resultatet tolkas som så att när variabeln 'Kön' ökar med 1 enhet ökar oddskvoten med ca 42 %. Resultatet påverkas även positivt av att föraren är yngre, där om vi ökar en enhet så minskar oddskvoten med 3 %. Variabeln 'Utbildare' har en ökning i oddskvoten på 152 % om man kör upp via trafikskola jämfört med om man kör upp som privatist. Ett annat viktigt resultat är att 'Befolk' och 'Stad' har en oddskvot nära 1, vilket innebär att de har väldigt svaga samband med resultatet på uppkörningen. De kategoriserade variablerna är aningen mer svårtolkade. Det vi kan konstatera är att för den mest nordliga latitud-kategorin (Lat_Kat100, ca Uppsala-Luleå) finns det en stor procentuell ökning av oddskvoten desto mer norr ut uppkörningen tar plats.

7 Diskussion

Om en person kör upp för första gången eller om det är ett omprov, är information som kan tänkas ha stor betydelse, men dessvärre inte tillgång till i datamaterialet. Detta innebär att en person som genomfört flera uppkörningar räknas som olika observationer. Detta gäller både omprov och uppgradering av behörigheten. Personer som innehaver andra former av körkort sedan tidigare skulle dessutom kunna ha bättre förutsättningar att klara uppkörningen, det gäller även B-körkort. Det skulle för framtida studier vara intressant att undersöka dessa faktorers inverkan på resultatet och huruvida de spelar en signifikant skillnad eller inte. Vi saknar

även uppgiften om till exempelvis väderlek vid körprovstillfället, modell på motorcykel och vilket moment individen misslyckats på.

Utöver tillgången på variabler behöver vi ifrågasätta vår modellframtagning och modellutvärdering. I purposeful selection strävas det efter att undersöka signifikanta kovariat, medan man i stegvis inkludering/exkludering vill ta fram den modell som ger det lägsta möjliga AIC:n. Enligt Hosmer-Lemeshow test är slutmodellen för purposeful selection den bästa, men studerar vi AIC:n så är slutmodellen för stepwise den bästa. När vi tar fram måttet på AUC:n visar det sig att modellen stepwise AIC med samspel verkar ha ett något bättre prediktionsmått, men skillnaderna är inte stora. Ingen av de slutgiltiga modellerna med samspel uppfyller kravet på $AUC > 0.7$, utan hamnar strax därunder (se tabell 11). Vi behöver observera att modeller med låg förklaringsgrad kan fortfarande ha god prediktionskraft. Dock, för att en modell ska anser vara bra behöver den både vara bra anpassad och ha god prediktionskraft.

Hosmer-Lemenshow är ett omdiskuterat test och användning av detta ska tas med försiktighet. Detta på grund av att testet inte tar hänsyn till överanpassning samt att det finns inte heller några tydliga riktlinjer kring vilket g man ska välja. Standard är g=10, men vilket g man väljer påverkar resultatet. Pearson χ^2 eller likelihood-ratio G^2 statistikan är bäst när de förklarande variablerna är enbart kategoriska (Agresti, 2013, s. 172). Om datan inte är grupperad så saknar χ^2 och G^2 gränsande χ^2 fördelningar. De kan dock fortfarande användas för att jämföra modeller enligt Agresti. På grund av detta ser vi till att inte enbart förlita oss på HL-testet utan även tar fram fler mått på anpassning och prediktionsförmåga. Desto fler tester som ger liknande indikation, ju mer tillförlitliga blir slutsatserna.

Hosmer och Lemeshow (2000, s. 164) skriver att de tror att ett riktigt värde för anpassning endast kan fås fram genom jämförelse mellan det observerade till predikterade värden från den anpassade modellen. Däremot kan det finnas situationer där R^2 kan användas för att ge nyttig statistik vid jämförelse av modeller som ska förklara samma data. Vi tar fram tre av de vanligaste R^2 måtten och samtliga mått indikerar på dålig anpassning för alla modeller och skillnaderna mellan modellerna är små.

Vi ser att modellerna med samspel ger oss något bättre AIC vilket är rimligt med tanke på att vi lägger till fler termer som gör att vi missar mindre information. Däremot är skillnaderna inte markanta och sålunda föredras en enklare modell framför en mer komplex. Ingen av modellerna är tillräckligt bra på att prediktera resultatet. Anledningarna till detta kan vara att vi saknar relevanta förklarande variabler i vår datamängd. Faktorer som skulle kunna tänkas vara av betydelse är till exempelvis antalet övningskörningstimmar, antal körda mil, väder på provdagen, om det är ett förstagångsprov eller inte och om föraren har bilkörkort eller inte.

Jämfört med Hjerts (2015) studie på sannolikheten att klara uppkörningen för bil fick vi lägre R^2-värden och sämre prediktionskraft. Hjert hade fler observationer i sitt datamaterial vilket kan vara en av anledningarna till bättre anpassning. AIC:n blev betydligt större i hans studie, vilket även kan förklaras av den stora datamängden. Gemensamt för Hjerts och denna studie är att 'Utbildare' och 'Kön' är två faktorer som uppenbart inverkar på resultatet av uppkörningen. Detta går även i hand med bland annat SMCs undersökningar. Vi finner, likt som Hjert, att en ökning av åldern tenderar att minska oddskvoten med ca 3 %.

8 Slutsats

De slutsatser vi kan dra från denna studie är att ingen av modellerna kan användas i predikativt syfte, däremot har analysen bidragit till att få ökad förståelse kring vilka faktorer som kan ha en inverkan på om en person klarar körprovet för motorcykel eller inte. Chanserna för att klara uppkörningen för motorcykel ökar om man är en ung man kör upp via körskola och högt upp i norr. Tillgång till fler betydande variabler är nödvändigt för att kunna ta fram en modell som predikterar resultatet.

9 Källförteckning

Agresti, A. (2013). Categorical data analysis - 3rd ed. John Wiley & Sons, Inc.

Held, L. & Bové, S., D. (2014). Applied statistical interference. Springer.

Hjert, T. (2015). Användning av logistisk regression vid bedömning av chansen att lyckas vid uppkörningen av B-körkort. Kand. Stockholms Universitet. Stockholm: Stockholms Universitet.

Hosmer Jr, D. W. & Lemeshow, S. (2000). Applied logistic regression - 2nd ed. John Wiley & Sons.

Mc-jakten. Handledare för mc-elev. Hemsida https://mc-jakten.se/mc-kort/handledare/ (Hämtat 2020-01-02).

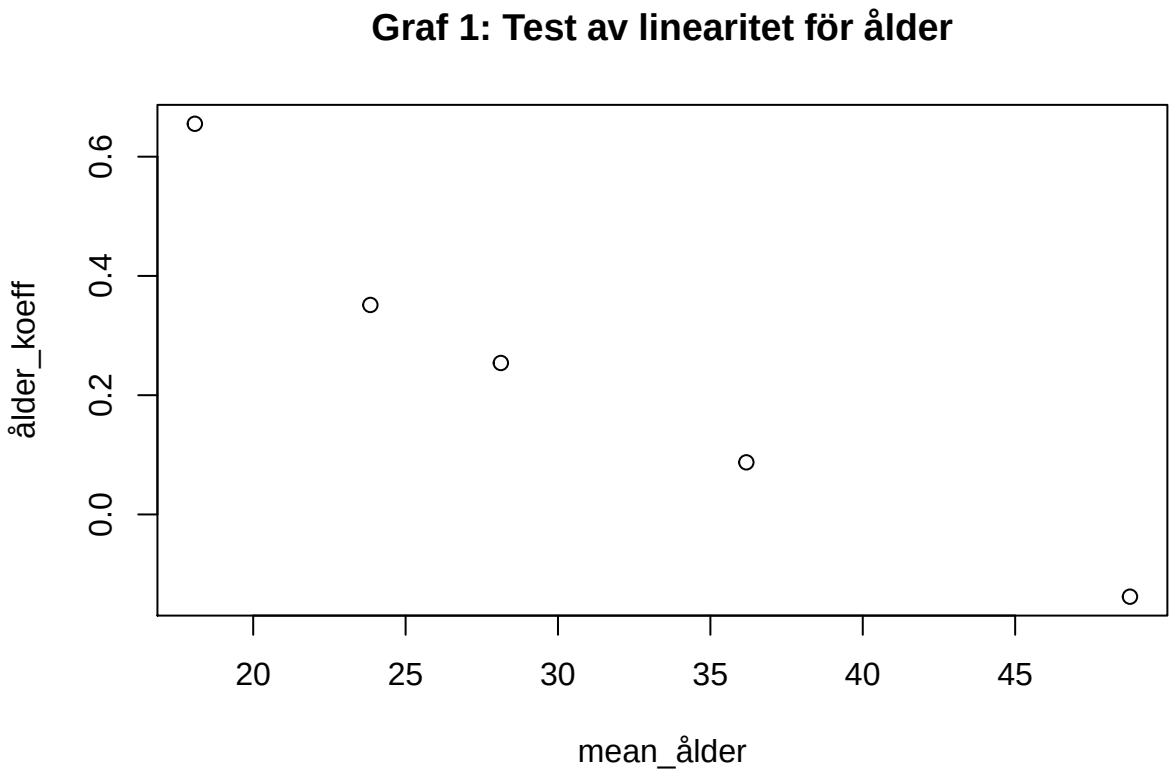
SMC. (2013). Körproven är inte rättvisa. https://www.svmc.se/smc/Nyheter/MC-forarnas-syn-pa-MC-prov-och-utbildning/ (Hämtat 2020-01-13).

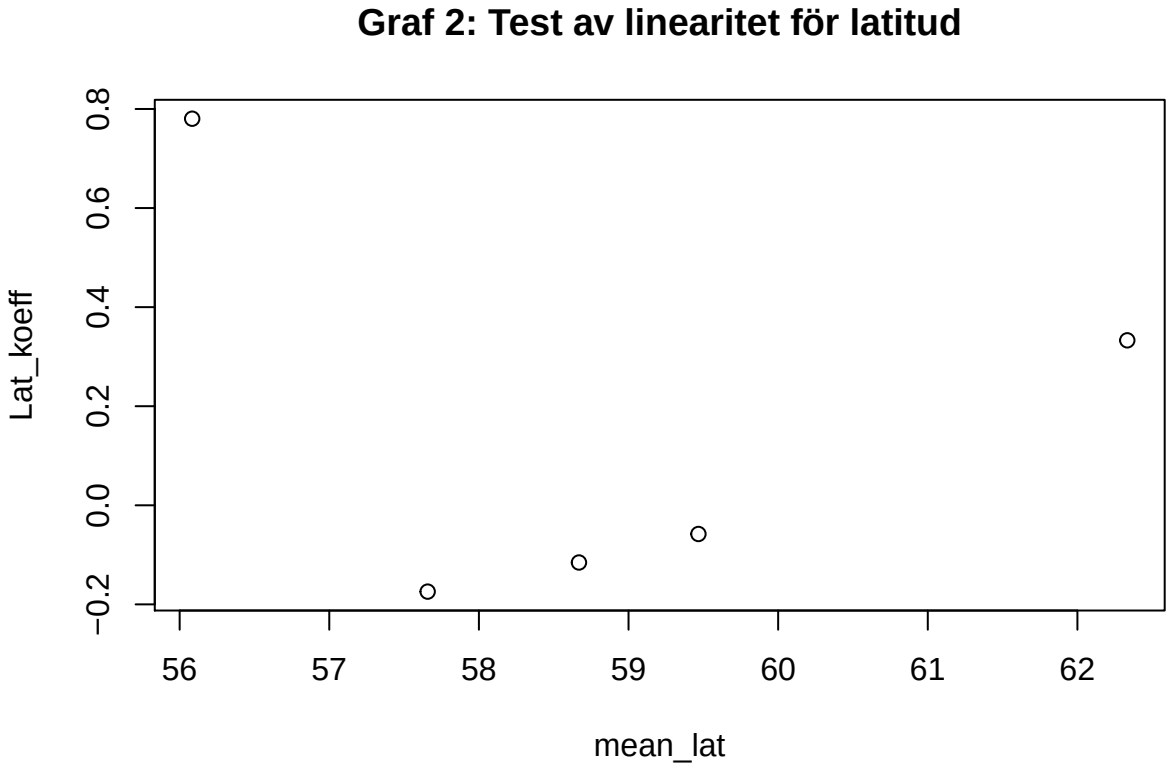
Sundberg, R. (2016). Lineära Statistiska Modeller.

Transportstyrelsen. (2015). Den 1 maj ändras manöverprovet för motorcykel. http://www.mynewsdesk.com/se/transportstyrelsen/pressreleases/den-1-maj-aendras-manoeverprovet-foer-motorcykel-1141787?utm_campaign=send_list&utm_medium=email&utm_source=sendgrid (Hämtat 2020-01-13).

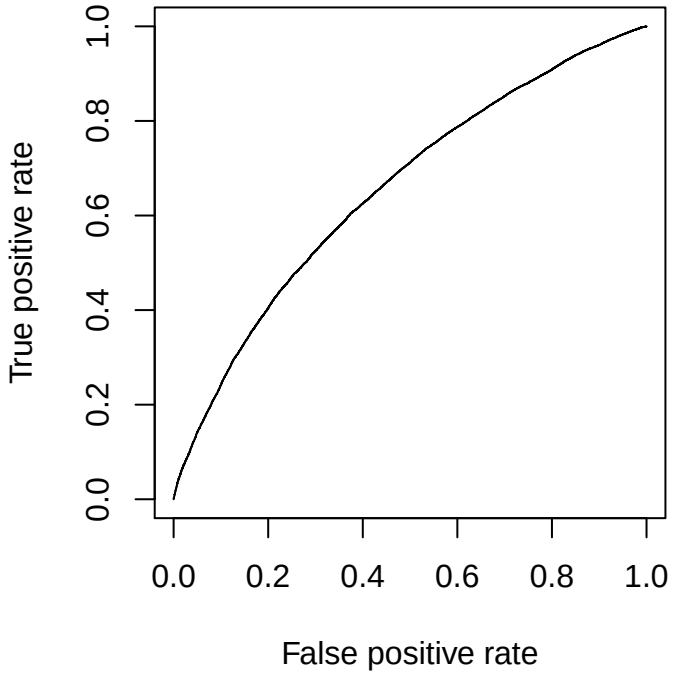
10 Bilaga

A Grafer

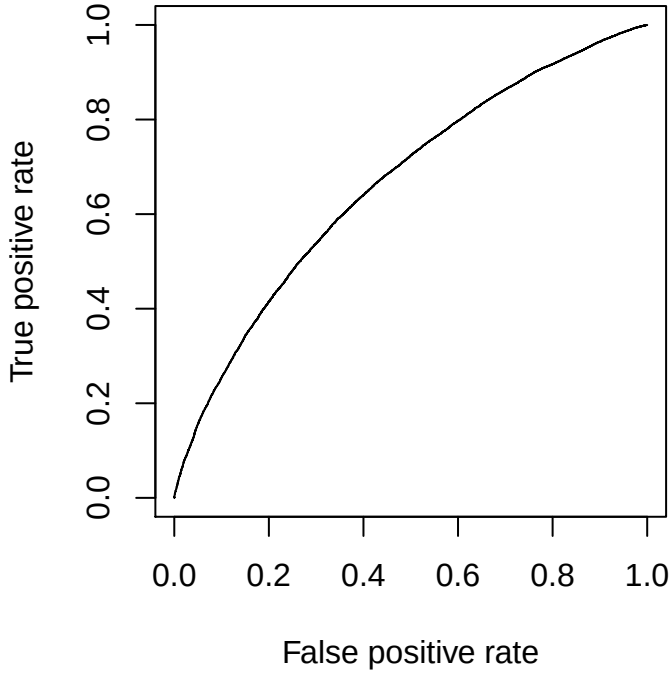




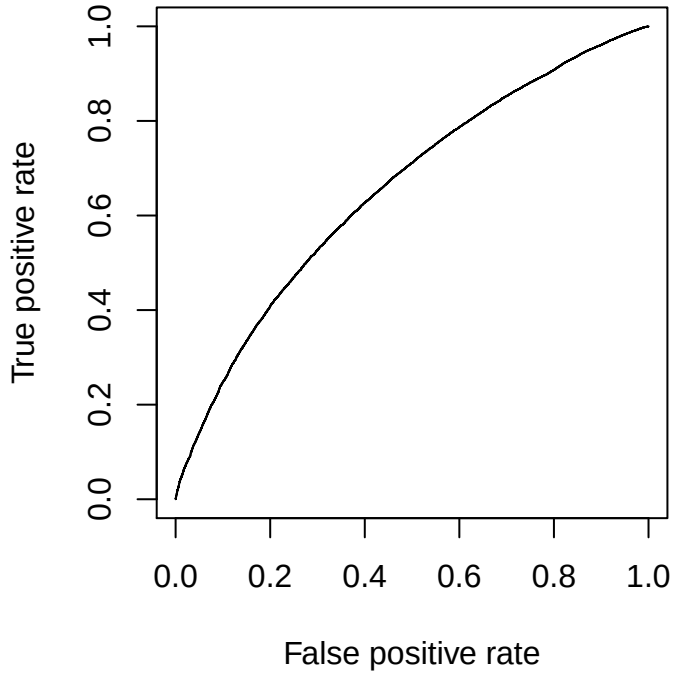
Graf 3: ROC kurva för Purposeful Selection utan samspel



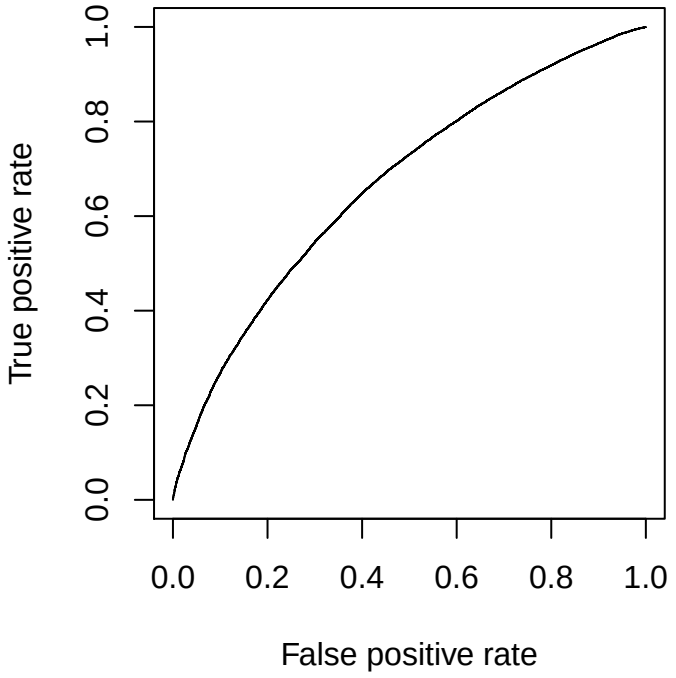
Graf 4: ROC kurva för Pursposeful Selection med samspel



Graf 5: ROC kurva för Stegvis AIC utan samspel



Graf 6: ROC kurva för Stegvis AIC med samspel



B Tabeller

Tabell 12: Signifikanta variabler från steg 1

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.4677	0.2965	-15.0685	0.0000
Kön	0.3504	0.0272	12.8922	0.0000
Ålder	-0.0265	0.0011	-24.4619	0.0000
ProvtypKörprov A1	-0.9076	0.0370	-24.5447	0.0000
ProvtypKörprov A2	-0.2811	0.0313	-8.9732	0.0000
Förmiddag	0.1025	0.0198	5.1662	0.0000
VeckodagLördag	-0.0104	0.0687	-0.1516	0.8795
VeckodagMåndag	0.1296	0.0402	3.2220	0.0013
VeckodagOnsdag	0.0824	0.0373	2.2093	0.0272
VeckodagSöndag	-0.0114	0.2251	-0.0505	0.9598
VeckodagTisdag	-0.0160	0.0364	-0.4392	0.6605
VeckodagTorsdag	0.1039	0.0370	2.8085	0.0050
SäsongSommar	0.0401	0.0225	1.7852	0.0742
SäsongVår	0.1182	0.0291	4.0612	0.0000
Utbildare	0.9311	0.0199	46.7844	0.0000
ProvplatsLat	0.0891	0.0049	18.1362	0.0000
Befolk	0.0000	0.0000	-0.2631	0.7925

Tabell 13: Purposeful Selection med samspel

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.0847	0.1346	8.0603	0.0000
Kön	-0.1912	0.0866	-2.2078	0.0273
Ålder	-0.0367	0.0034	-10.8898	0.0000
ProvtypKörprov A1	-0.7311	0.1143	-6.3982	0.0000
ProvtypKörprov A2	-0.4533	0.0946	-4.7917	0.0000
Förmiddag	0.2529	0.0713	3.5488	0.0004
VeckodagLördag	-0.2387	0.0971	-2.4579	0.0140
VeckodagMåndag	0.0722	0.0609	1.1844	0.2362
VeckodagOnsdag	0.0513	0.0577	0.8887	0.3741
VeckodagSöndag	0.0408	0.3231	0.1262	0.8996
VeckodagTisdag	-0.1114	0.0555	-2.0066	0.0448
VeckodagTorsdag	-0.0318	0.0563	-0.5652	0.5719
SäsongSommar	-0.0214	0.0331	-0.6472	0.5175
SäsongVår	0.0841	0.0409	2.0580	0.0396
Utbildare	0.9400	0.1256	7.4846	0.0000
Lat_Kat40	-0.1158	0.0977	-1.1858	0.2357
Lat_Kat60	0.1087	0.0986	1.1024	0.2703
Lat_Kat80	-0.1242	0.0997	-1.2456	0.2129
Lat_Kat100	-0.0126	0.1130	-0.1116	0.9112
Stad	0.0000	0.0000	-3.5170	0.0004
Kön:Ålder	0.0124	0.0023	5.3680	0.0000
Kön:Utbildare	0.2327	0.0557	4.1760	0.0000
Ålder:Förmiddag	-0.0048	0.0018	-2.7151	0.0066
Ålder:Utbildare	-0.0059	0.0023	-2.6138	0.0090
Ålder:Lat_Kat40	0.0043	0.0027	1.6007	0.1095
Ålder:Lat_Kat60	0.0098	0.0027	3.6547	0.0003

Ålder:Lat_Kat80	0.0038	0.0027	1.4209	0.1553
Ålder:Lat_Kat100	0.0132	0.0030	4.4877	0.0000
ProvtypKörprov A1:Utbildare	-0.2634	0.0753	-3.4996	0.0005
ProvtypKörprov A2:Utbildare	-0.0173	0.0643	-0.2690	0.7879
Förmiddag:Lat_Kat40	-0.0451	0.0616	-0.7320	0.4642
Förmiddag:Lat_Kat60	-0.1594	0.0616	-2.5894	0.0096
Förmiddag:Lat_Kat80	0.0156	0.0613	0.2547	0.7990
Förmiddag:Lat_Kat100	0.2630	0.0775	3.3935	0.0007
SäsongSommar:Utbildare	0.1246	0.0454	2.7443	0.0061
SäsongVår:Utbildare	0.0406	0.0589	0.6903	0.4900
Utbildare:Lat_Kat40	-0.0822	0.0627	-1.3097	0.1903
Utbildare:Lat_Kat60	-0.3631	0.0629	-5.7696	0.0000
Utbildare:Lat_Kat80	-0.2810	0.0634	-4.4326	0.0000
Utbildare:Lat_Kat100	0.0245	0.0703	0.3490	0.7271
Utbildare:Stad	0.0000	0.0000	2.4483	0.0144
Lat_Kat40:Stad	0.0000	0.0000	5.0098	0.0000
Lat_Kat60:Stad	0.0584	0.0535	1.0917	0.2750
Lat_Kat80:Stad	0.6124	0.0418	14.6629	0.0000
Lat_Kat100:Stad	0.0000	0.0000	1.0269	0.3045
ProvtypKörprov A1:VeckodagLördag	0.2228	0.2120	1.0508	0.2934
ProvtypKörprov A2:VeckodagLördag	0.4361	0.2066	2.1106	0.0348
ProvtypKörprov A1:VeckodagMåndag	-0.0248	0.1365	-0.1815	0.8560
ProvtypKörprov A2:VeckodagMåndag	0.2505	0.1155	2.1694	0.0301
ProvtypKörprov A1:VeckodagOnsdag	-0.1383	0.1250	-1.1064	0.2685
ProvtypKörprov A2:VeckodagOnsdag	0.2468	0.1070	2.3073	0.0210
ProvtypKörprov A1:VeckodagSöndag	-1.6723	0.6586	-2.5391	0.0111
ProvtypKörprov A2:VeckodagSöndag	0.5502	0.8980	0.6127	0.5401
ProvtypKörprov A1:VeckodagTisdag	-0.0533	0.1233	-0.4326	0.6653
ProvtypKörprov A2:VeckodagTisdag	0.1539	0.1038	1.4834	0.1380
ProvtypKörprov A1:VeckodagTorsdag	0.0144	0.1247	0.1153	0.9082
ProvtypKörprov A2:VeckodagTorsdag	0.1868	0.1059	1.7638	0.0778
VeckodagLördag:Utbildare	0.4391	0.1523	2.8833	0.0039
VeckodagMåndag:Utbildare	-0.0010	0.0818	-0.0122	0.9903
VeckodagOnsdag:Utbildare	-0.0817	0.0762	-1.0728	0.2834
VeckodagSöndag:Utbildare	0.6662	0.5828	1.1431	0.2530
VeckodagTisdag:Utbildare	0.1355	0.0741	1.8296	0.0673
VeckodagTorsdag:Utbildare	0.1365	0.0751	1.8187	0.0690

Tabell 14: Stegvis AIC med samspel

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.4862	0.2313	6.4241	0.0000
Kön	-0.0313	0.1157	-0.2702	0.7870
Ålder	-0.0320	0.0037	-8.6194	0.0000
ProvtypKörprov A1	-0.4599	0.3468	-1.3261	0.1848
ProvtypKörprov A2	0.3302	0.2167	1.5237	0.1276
Förmiddag	0.1235	0.0883	1.3977	0.1622
VeckodagLördag	-0.2960	0.5603	-0.5283	0.5973
VeckodagMåndag	-0.1656	0.1858	-0.8913	0.3727
VeckodagOnsdag	-0.5157	0.1846	-2.7937	0.0052
VeckodagSöndag	32.5054	30.3721	1.0702	0.2845

VeckodagTisdag	-0.4469	0.1814	-2.4635	0.0138
VeckodagTorsdag	-0.2539	0.1832	-1.3857	0.1658
SäsongSommar	-0.0272	0.0591	-0.4607	0.6450
SäsongVår	0.0399	0.0750	0.5317	0.5949
Utbildare	1.0358	0.1315	7.8795	0.0000
Lat_Kat40	-0.7607	0.2083	-3.6525	0.0003
Lat_Kat60	-0.2733	0.2210	-1.2370	0.2161
Lat_Kat80	-0.6351	0.1910	-3.3260	0.0009
Lat_Kat100	-0.5185	0.2301	-2.2535	0.0242
Befolk	0.0000	0.0000	-3.7496	0.0002
Stad	0.0001	0.0001	0.8203	0.4120
Kön:Ålder	0.0091	0.0029	3.1779	0.0015
Kön:ProvtypKörprov A1	-0.1663	0.1270	-1.3091	0.1905
Kön:ProvtypKörprov A2	-0.2162	0.0993	-2.1766	0.0295
Kön:Utbildare	0.2215	0.0560	3.9553	0.0001
Ålder:ProvtypKörprov A1	-0.0098	0.0185	-0.5282	0.5973
Ålder:ProvtypKörprov A2	-0.0195	0.0087	-2.2290	0.0258
Ålder:Förmiddag	-0.0045	0.0018	-2.5303	0.0114
Ålder:Utbildare	-0.0068	0.0023	-2.9837	0.0028
Ålder:Lat_Kat40	0.0031	0.0027	1.1624	0.2451
Ålder:Lat_Kat60	0.0083	0.0027	3.0823	0.0021
Ålder:Lat_Kat80	0.0023	0.0027	0.8730	0.3827
Ålder:Lat_Kat100	0.0118	0.0030	3.9829	0.0001
ProvtypKörprov A1:Utbildare	-0.2855	0.0757	-3.7710	0.0002
ProvtypKörprov A2:Utbildare	-0.0309	0.0645	-0.4793	0.6317
Förmiddag:Utbildare	-0.0462	0.0427	-1.0815	0.2795
Förmiddag:Lat_Kat40	0.1059	0.0758	1.3963	0.1626
Förmiddag:Lat_Kat60	-0.0600	0.0746	-0.8033	0.4218
Förmiddag:Lat_Kat80	0.1193	0.0693	1.7206	0.0853
Förmiddag:Lat_Kat100	0.2207	0.0886	2.4900	0.0128
SäsongSommar:Utbildare	0.1171	0.0469	2.4984	0.0125
SäsongVår:Utbildare	0.0337	0.0608	0.5544	0.5793
SäsongSommar:Lat_Kat40	-0.0601	0.0690	-0.8719	0.3833
SäsongVår:Lat_Kat40	0.1255	0.0892	1.4061	0.1597
SäsongSommar:Lat_Kat60	0.0529	0.0704	0.7511	0.4526
SäsongVår:Lat_Kat60	0.0358	0.0892	0.4017	0.6879
SäsongSommar:Lat_Kat80	0.0295	0.0702	0.4202	0.6744
SäsongVår:Lat_Kat80	0.0956	0.0901	1.0609	0.2887
SäsongSommar:Lat_Kat100	-0.0687	0.0806	-0.8526	0.3939
SäsongVår:Lat_Kat100	-0.1262	0.1068	-1.1816	0.2374
Utbildare:Lat_Kat40	-0.1220	0.0647	-1.8869	0.0592
Utbildare:Lat_Kat60	-0.4258	0.0644	-6.6086	0.0000
Utbildare:Lat_Kat80	-0.3339	0.0647	-5.1575	0.0000
Utbildare:Lat_Kat100	-0.0553	0.0727	-0.7603	0.4471
Förmiddag:Befolk	0.0000	0.0000	1.5146	0.1299
Lat_Kat40:Befolk	0.0000	0.0000	7.6249	0.0000
Lat_Kat60:Befolk	0.0000	0.0000	-2.5272	0.0115
Lat_Kat80:Befolk	0.0000	0.0000	1.6025	0.1090
Lat_Kat100:Befolk	0.0000	0.0000	-3.6236	0.0003
Utbildare:Stad	0.0000	0.0000	2.4286	0.0152
Lat_Kat40:Stad	0.0000	0.0000	3.2357	0.0012

Lat_Kat60:Stad	0.6486	0.1898	3.4164	0.0006
Lat_Kat80:Stad	0.6188	0.0815	7.5959	0.0000
Lat_Kat100:Stad	0.0000	0.0000	0.4580	0.6470
VeckodagLördag:Utbildare	0.5396	0.1598	3.3772	0.0007
VeckodagMåndag:Utbildare	0.0557	0.0842	0.6608	0.5087
VeckodagOnsdag:Utbildare	-0.0206	0.0783	-0.2625	0.7929
VeckodagSöndag:Utbildare	-0.0691	0.8588	-0.0805	0.9359
VeckodagTisdag:Utbildare	0.1709	0.0757	2.2559	0.0241
VeckodagTorsdag:Utbildare	0.1638	0.0767	2.1347	0.0328
VeckodagLördag:Lat_Kat40	-0.0694	0.5669	-0.1223	0.9026
VeckodagMåndag:Lat_Kat40	0.0985	0.1916	0.5144	0.6070
VeckodagOnsdag:Lat_Kat40	0.6376	0.1870	3.4106	0.0006
VeckodagSöndag:Lat_Kat40	-30.6139	30.2105	-1.0134	0.3109
VeckodagTisdag:Lat_Kat40	0.2881	0.1840	1.5655	0.1175
VeckodagTorsdag:Lat_Kat40	0.1697	0.1846	0.9195	0.3578
VeckodagLördag:Lat_Kat60	0.0871	0.5314	0.1639	0.8698
VeckodagMåndag:Lat_Kat60	0.1709	0.2004	0.8526	0.3939
VeckodagOnsdag:Lat_Kat60	0.4721	0.1945	2.4273	0.0152
VeckodagSöndag:Lat_Kat60	-30.2999	27.6766	-1.0948	0.2736
VeckodagTisdag:Lat_Kat60	0.3250	0.1924	1.6896	0.0911
VeckodagTorsdag:Lat_Kat60	0.2412	0.1952	1.2354	0.2167
VeckodagLördag:Lat_Kat80	0.3666	0.5173	0.7086	0.4786
VeckodagMåndag:Lat_Kat80	0.2037	0.1638	1.2440	0.2135
VeckodagOnsdag:Lat_Kat80	0.4626	0.1607	2.8792	0.0040
VeckodagTisdag:Lat_Kat80	0.1955	0.1571	1.2449	0.2132
VeckodagTorsdag:Lat_Kat80	0.1760	0.1586	1.1099	0.2671
VeckodagLördag:Lat_Kat100	0.8082	0.5613	1.4398	0.1499
VeckodagMåndag:Lat_Kat100	0.6905	0.2062	3.3490	0.0008
VeckodagOnsdag:Lat_Kat100	0.7983	0.2022	3.9489	0.0001
VeckodagSöndag:Lat_Kat100	-21.6505	55.2493	-0.3919	0.6952
VeckodagTisdag:Lat_Kat100	0.6697	0.1987	3.3702	0.0008
VeckodagTorsdag:Lat_Kat100	0.5923	0.2003	2.9573	0.0031
VeckodagLördag:Befolk	0.0000	0.0000	-0.7143	0.4750
VeckodagMåndag:Befolk	0.0000	0.0000	0.0412	0.9671
VeckodagOnsdag:Befolk	0.0000	0.0000	1.9118	0.0559
VeckodagSöndag:Befolk	-0.0001	0.0001	-0.9888	0.3228
VeckodagTisdag:Befolk	0.0000	0.0000	1.3619	0.1732
VeckodagTorsdag:Befolk	0.0000	0.0000	0.1554	0.8765
VeckodagLördag:Stad	0.1468	0.2949	0.4980	0.6185
VeckodagMåndag:Stad	-0.0001	0.0001	-1.5893	0.1120
VeckodagOnsdag:Stad	-0.0001	0.0001	-0.9816	0.3263
VeckodagSöndag:Stad	6.8126	9.0186	0.7554	0.4500
VeckodagTisdag:Stad	-0.0001	0.0001	-1.4057	0.1598
VeckodagTorsdag:Stad	-0.0001	0.0001	-1.2009	0.2298

Tabell 15: Oddskvot Stegvis AIC

Modell	Oddskvot
Kön	1.422
Ålder	0.974
ProvtypKörprov A1	0.401
ProvtypKörprov A2	0.753
Förmiddag	1.065
VeckodagLördag	1.003
VeckodagMåndag	1.135
VeckodagOnsdag	1.091
VeckodagSöndag	1.013
VeckodagTisdag	0.983
VeckodagTorsdag	1.098
SäsongSommar	1.038
SäsongVår	1.121
Utbildare	2.522
Lat_Kat40	0.884
Lat_Kat60	1.088
Lat_Kat80	1.199
Lat_Kat100	1.707
Befolk	0.999
Stad	1.000