Stockholms universitet

Statistisk analys av faktorer som påverkar studieframgången inom kursen matematik I på Stockholms universitet

Malin Andersson

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2015:10
Matematisk statistik
Juni 2015

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2015:10**
http://www.math.su.se/matstat

Statistisk analys av faktorer som påverkar studieframgången inom kursen matematik I på Stockholms universitet

Malin Andersson*

Juni 2015

Sammanfattning

Cirka 4000 studenter var registrerade vid kursen Matematik I mellan höstterminen 2007 och vårterminen 2014. I denna uppsats analyserades 1547 av dessa studenterna. Studenterna vi analyserar är tillhörande ett kandidatprogram och läser kursen på helfart. Studenterna studerar på olika kandidatprogram, de har olika syften och mål med sina studier. De har även olika förutsättningar och bakgrund vilket leder till att de presterar olika. Syftet med uppsatsen var att undersöka vilka faktorer som påverkar och spelar en viktig roll för huruvida en student som registrerats på kursen matematik I klarar av kursen samma termin som man registrerades för första gången. Detta har kommit att bli en intressant fråga då samtidigt som studentkullarna ökar så ökar även eftersläpande studenter, vilket kan leda till att man har fler studenter än vad resurserna tillåter. På kursen matematik I har omregistreringsrutinerna för studenter som inte klarade kursen samma termin ändrats från och med vårterminen 2015. Man har inte längre möjlighet att garantera att alla får en omregistrering. Vi har använt oss av logistisk regression för att plocka fram den modell som i detta fall kan förklara mest om huruvida en student klarar kursen samma termin eller vid ett senare tillfälle under perioden 2007-2014. Ålder vid registrering har kommit att bli den faktor som är mest avgörande för huruvida studenten klarar kursen samma termin, där yngre studenter klarar sig bättre än äldre. Det har även visat sig att andelen studenter som klarar kursen samma termin har minskat med åren vilket således innebär att andelen eftersläpande studenter ökar. Studenter klarar kursen allt mer sällan samma termin som man registrerades på den för första gången. Det har även visat sig råda en skillnad i prestationer mellan de olika kandidatprogrammen.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: malinandersson10@hotmail.com. Handledare: Martin Sköld & Jan-Olov Persson.

Förord

Denna kandidatuppsats i matematisk statistik är skriven vid matematiska institutionen på Stockholms universitet. Arbetet omfattar 15 högskolepoäng och leder till en kandidatexamen i matematisk statistik.

Jag vill rikta ett stort tack till mina handledare Martin Sköld och Jan-Olov Persson som bidragit med värdefull handledning och som genom hela arbetet visat stort engagemang. Jag vill även tacka Filip Walldén, min kurskamrat som bidragit med värdefulla synpunkter. Till sist vill jag även tacka min sambo John Phiri och min mamma Lena Andersson för deras stöd genom hela utbildningen.

Innehållsförteckning

1. Introduktion

1.1 Matematik I

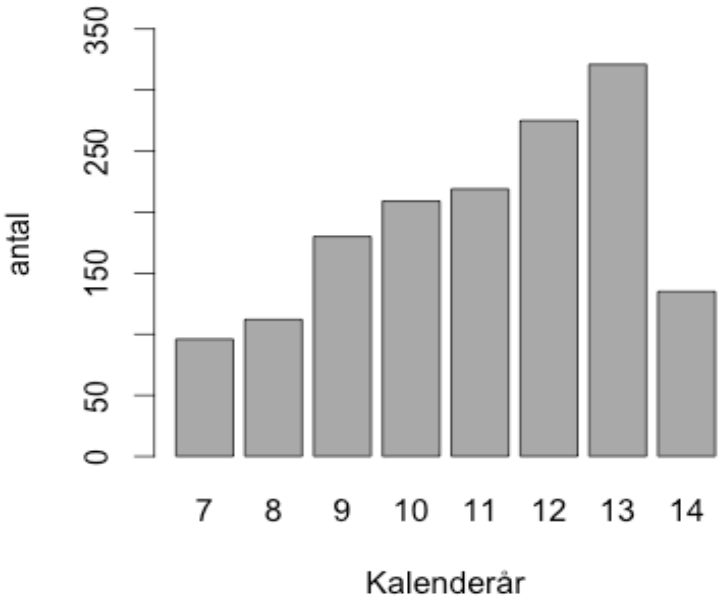
Matematik I är en grundkurs i matematik som ges vid Stockholms Universitet. Kursen innefattar grundläggande algebra, funktionslära, linjär algebra, envariabelanalys och flervariabelanalys. För många studenter är de den första matematikkursen man läser på universitetsnivå och för många den första kursen på universitetsnivå överhuvudtaget. Kursen ges varje termin och kan läsas som en del av ett kandidatprogram men även som fristående kurs. Kursen innefattar 30 högskolepoäng, där man läser 15 högskolepoäng i algebra respektive analys. Undervisningen innefattar föreläsningar, handledningar samt räkneövningar.
Under kursens gång har man även möjlighet att samla bonuspoäng med hjälp av problemsamlingar till den skriftliga sluttentamen. Betygsskalan sträcker sig från A-E där man får ett sammanvägt betyg för de båda delmomenten algebra och analys. De två sluttentorna ger studenterna totalt 15 högskolepoäng och resterande 15 poängen samlas ihop genom obligatoriska e-tentor, seminarium och datorlaborationer.
För att ha klarat av hela kursen och kunna tillgodoräkna sig den krävs det att man avklarat alla 30 högskolepoängen.
Studenterna kan välja att läsa kursen på antingen halv eller helfart. Detta gäller enbart fristående kursare då alla kandidatprogram exklusive datalogi studenterna läser kursen på helfart enligt kursplanen. Kursen kan läsas på campus eller distans. (Matematiska, 2015)
För den nyfikna läsaren finns mer information om kursupplägget på matematiska institutionens hemsida.

1.2 Bakgrund

Genomströmningen vid högre studier är en viktig fråga för de flesta. Både samhället, universitetet, institutionen och individen själv. SACO skriver att efter många år av ökade studentkullar i kombination med att pengar till lärosätena inte ökat i samma takt, pressas resurserna på universitet och högskolor. Vilket i sig kan leda till ett ökat eftersläp. (Ehlin Kolk, 2012)

Sett till matematiska institutionen och kursen matematik I är söktrycket så pass högt att omregistreringsrutinerna har ändrats från och med vårterminen 2015. Man har inte längre möjlighet att garantera att alla får en omregistrering. Det blir med andra ord viktigare att studenterna klarar kursen samma termin som man registrerades på den för första gången. I figur 1 nedan ser vi hur antal registreringar ökat sedan 2007 fram till 2013. Anledningen till nergången under 2014 beror på att vi enbart har data för vårterminen det kalenderåret. För övriga år innefattar kalenderåret både höst och vårtermin. (Matematiska, 2015)

Figur 1, Antal registreringar på kursen matematik 1



1.3 Syfte

Syftet med denna uppsats är att göra en statistisk analys av vilka faktorer som kan förklara studieframgången hos studenterna som klarar kursen samma termin som man registrerades för första gången. De studenter som inte blir godkända samma termin kommer rapportera ett godkänt betyg senare under perioden 2007-2014. Anledningen till att vi valt att begränsa oss till enbart de studenter som antingen får ett godkänt betyg samma termin eller inom perioden vi tittar på har och göra med att de är många registrerade studenter som aldrig dyker upp eller slutför kursen. Vi vet således inte vad deras plan är, ska de ta kursen vid ett senare tillfälle? Blev de antagna till något annat program? Lämnade de institutionen helt och hållet?

Studenterna registrerar sig och dyker helt enkelt sedan inte upp av olika anledningar.Huvudsakligen kommer logistisk regression med binär responsvariabel att användas.

Y=0: Studenten får inte ett godkänt betyg samma termin utan senare under perioden.

Y=1: Studenten får ett godkänt betyg samma termin

Programvaran R har använts genom hela arbetet.

1.4 Frågeställningar

Vilka faktorer är de som påverkar och spelar en viktig roll för huruvida en student som registrerats på kursen matematik I klarar av kursen samma termin som man registrerades för första gången? Givet att de studenter som inte klarade sig samma termin får ett godkänt betyg senare under perioden 2007-2014. Detta begränsat till variabler som kunnat samlats ihop i Ladok.

2. Beskrivning av data

Data som finns tillgänglig har samlats in under våren 2015. I datamaterialet finns antalet godkända helfarts studenter som varit registrerade på kursen matematik I vid någon termin från höstterminen 2007 fram till vårterminen 2014.
Vi har begränsat oss till enbart heltfarts studenter då vi ifrån Ladok inte har kunnat hämta uppgifter om huruvida studenter läser kursen på hel eller halvfart. Vi vet att alla studenter tillhörande ett kandidatprogram läser kursen på helfart exklusive datalogerna som läser algebradelen under år 1 och analysdelen under år 2. Datalogi studenterna blev därav exkluderade från materialet. Vi valde även att exkludera de fristående kursarna. Anledningen till att de plockades bort var att de kunde leda till ett missvisade resultat om vi antog att majoriteten läste på helfart och det visade sig att majoriteten faktiskt läste på halvfart. Om man då lät alla studenter bara ha en termin på sig att klara kursen då de egentligen skulle haft ett helt år skulle de se ut som att fler fristående kursare bidrog till eftersläpet än vad den faktiska siffran är.

Som vi nämnde tidigare har vi begränsat oss till studenter som någon gång under perioden 2007-2014 rapporterat in ett godkänt betyg på kursen. Detta då det kan bli missvisande att titta på alla studenter som varit registrerade på kursen eftersom de är känt att studenter registrerar sig och sedan inte har för avsikt att avsluta den. Vad som är viktigt att notera då man gör en sådan begränsning är att studenter som är registrerade längre bak i tiden, låt oss säga 2007 har haft längre tid på sig att rapportera in ett godkänt betyg. Förhoppningsvis har de som registrerade sig kalenderåret 2007 rapporterat in ett godkänt betyg om man hade för avsikt att slutföra kursen. Vilket inte är lika troligt för studenter som registrerades 2014 då vi inte har uppgifter för huruvida de rapporterar in ett godkänt betyg senare än den terminen de registrerades på.

I Datamaterialet har vi uppgifter om studenterna som innefattar; kön, ålder vid registrering, program, betyg, kalenderåret man blev registrerad, om man började höst- eller vårtermin samt huruvida man valt att läsa en förberedande kurs i matematik som institutionen erbjuder. I tabell 1 nedan kan vi se hur studenterna fördelar sig mellan de olika kovariaterna.

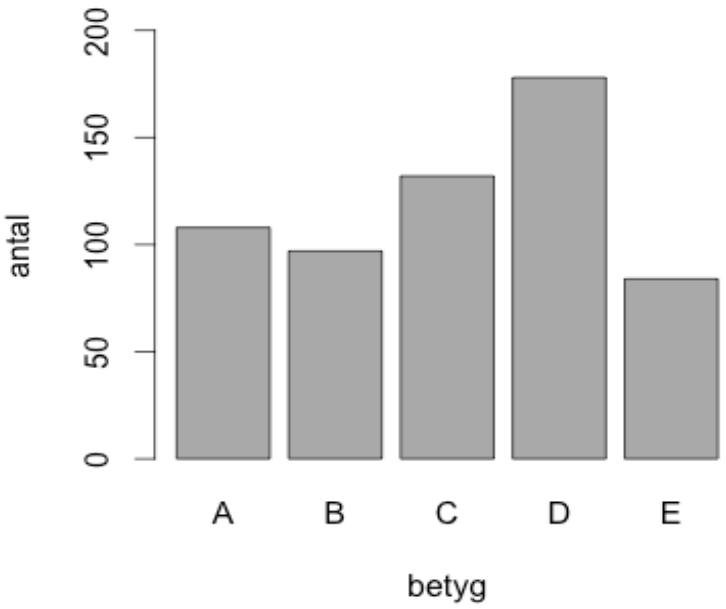
Tabell 1, Förklaring av kovariater, procenten grovt avrundade

Kovariater		**Antal(procent)**
Kön	Kvinna	300 (50)
	Man	299 (50)
Terminsstart	Höst	469 (78)
	Vår	130 (22)
Program	Lärare	41 (7)
	Astronomi	46 (8)
	Biofysik	6 (1)
	Biomatematik och beräkningsbiologi	29 (5)
	Fysik	109 (18)
	Matematik	151 (25)
	Meteorologi	41 (7)
	Matematik och filosofi	22 (4)

	Sjukhusfysiker	55(9)
	Matematik och ekonomi	99 (16)
Totalt		**599**

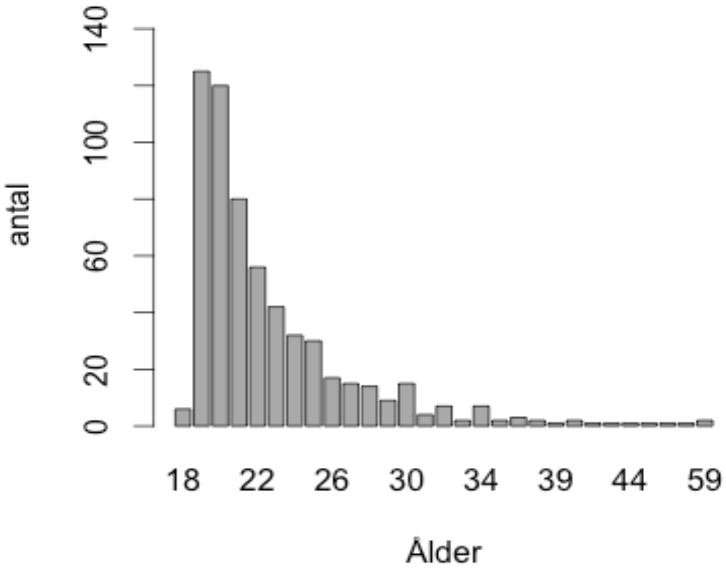
Antalet registreringar på kursen mellan höstterminen 2007 och vårterminen 2014 är 1547 stycken. Av dessa var det 599 studenter som fått ett godkänt betyg. För att få ett godkänt betyg i kursen krävs det att studenterna klarat av alla momenten. Vi gör således inte skillnad på analys och algebra. Betygen motsvarar det sammanvägda betyget från algebra och analys. I nedanstående figur 2 kan vi se betygsspridningen för de studenter som fått ett godkänt betyg på kursen.

Figur 2, betygsfördelning för godkända studenter



Könsfördelningen är jämt fördelad med 51 % män och 49 % kvinnor.
I figur 3 kan vi se åldersspridningen över de studenter som någon gång under perioden fick ett godkänt betyg. Yngsta registrerade studenten vid kursstart är 18 år och äldsta 59 år.

Figur 3, Åldersfördelningen bland de 599 studenterna



3. Metoder

I denna del kommer vi presentera de statistiska metoder som använts för att analysera huruvida en student klarar sig samma termin eller senare under perioden 2007-2014.

3.1 Logistisk regression

Regressionsmetoder har kommit att bli en naturlig del av en analys då man vill undersöka ett förhållande mellan en responsvariabel (y) och en eller flera förklarande variabler (x). Förklarande variablerna kallas ofta för kovariter. Målet med denna typ av metod är att hitta den bäst passande modellen som beskriver förhållandet mellan responsvariabeln och en eller flera kovariater. Med den ”bästa” modellen menar vi den modell som bäst passar de ändamål vi har med analysen. (W.Hosmer & Lemeshow, 2000, s.1) I detta fall har vi en binär responsvariabel *(y)* samt kovariater *(x).* Då man har att göra med en binär responsvariabel finns det två möjliga utfall av ett försök. Med andra ord kan responsvariabeln anta två värde, *Y=1* samt *Y=0.* I detta fall vill vi beräkna sannolikheten att *Y=1* beroende på vilka värden x kommer att anta. Kort sagt vill vi undersöka vad som händer med Y då x ökar eller minskar.

$\pi(x) = P(Y = 1|X = x)$, *Sannolikheten att utfallet inträffar*
$\pi(x) = 1 - P(Y = 0|X = x)$, *Sannolikheten att utfallet inte inträffar*

I regel då vi har att göra med en responsvariabel av detta slag har vi ett icke-linjärt samband mellan $\pi(x)$ och x. Det visar sig även att sambandet mellan $\pi(x)$ och x ofta är monotont (växande eller avtagande) där $\pi(x)$ ökar eller minskar då x ökar.

Den logistiska regressions modellen ges av: $\pi(x) = \frac{exp(\alpha+\beta x)}{1+exp(\alpha+\beta x)}$

Viktigt att notera att då x ökar kommer $\pi(x)$ öka om $\beta > 0$ och minska då $\beta < 0$
Ekvivalent med logistiska regressions modellen så har logit (log oddset) det linjära förhållandet:

$logit[\pi(x)] = \log\frac{\pi(x)}{1-\pi(x)} = \alpha + \beta x$

Det är detta odds vi studerar för att ta reda på oddset för att Y=1 ska inträffa.
Har man att göra med flera kovariater byts βx mot $(\beta_1 x_1 + \ldots + \beta_p x_p)$
(Agresti, 2013, s. 163-164, 119-120)

3.1.1 Log-odds & odds kvot

Odds är ett mått på hur sannolikt det är att en viss händelse inträffar respektive att den inte inträffar.

Lyckat försök: π

Misslyckat försök: 1- π

Oddset definieras som:

$\frac{Sannolikheten\ av\ lyckat\ försök}{Sannolikheten\ av\ misslyckat\ förösk} = \frac{\pi}{1-\pi} = exp(\alpha + \beta_1 x_1 + .. + \beta_p x_p)$

Oddset ligger alltid mellan 0 och ∞, där ett odds>1 innebär att ett lyckat försök är mer troligt än ett misslyckat. Om oddset ligger mellan 0 och 1, dvs 0<odds<1, är misslyckat försök mest troligt. Om oddset=1 är sannolikheten för misslyckat respektive lyckat försök lika troligt.

Oddskvoten är kvoten mellan oddset för händelsen Y=1 (I detta fall att studenten får ett godkänt betyg samma termin) om x_1 ökar med en enhet och oddset för Y=1 om vi håller x_1 konstant. När vi säger att x_1 ökar med en enhet skulle det i vårt fall exempelvis kunna vara att åldern ökar med ett år. Detta är vad som gäller för kontinuerliga variablerna. För kategoriska variabler är oddskvoten kvoten mellan oddset för Y=1 inom en kategori av x jämfört med oddset för att Y=1 i referenskategorin.

Odds-kvoten definieras då som: $Oddskvoten = \frac{exp\left(\alpha+\beta_1 x_1+\ldots+\beta_p(x_p+1)\right)}{exp(\alpha+\beta_1 x_1+\ldots+\beta_p x_p)} = exp\ (\beta_p)$

(Agresti, 2013, s. 44-45, s.164)

Log-oddset är då vi logaritmerar både höger och vänster led av oddset.

Vilket ges av: $log = \frac{\pi(x)}{1-\pi(x)} = \alpha + \beta_1 x_1 + .. + \beta_p x_p$

(Ibid, 2013, s.119)

3.2 Multikollinearitet

Även om vi i denna modell har relativt få potentiella förklarandevariabler kan det ibland förekomma linjära samband mellan dem. Multikollinearitet är ett linjärt samband mellan en eller flera variabler. Har man variabler i modellen som innefattar liknande information kan det innebära att de är korrelerade. Således kan problem med multikollineatitet uppstå. (Sundberg, 2014, s.71) Korrelationen mellan variablerna kan få det att framstå som att ingen av variablerna är viktiga. En variabel kan även se ut att ha en liten effekt för att den ’överlappas’ av de andra kovariaterna. Om så skulle vara fallet, kan det vara hjälpsamt att exkludera denna variabel. (Agresti, 2013, s.208)

3.3 Stegvisa procedurer

Med ett antal olika förklarande variabler finns det olika potentiella modeller.

De stegvisa procedurer vi använt oss av i detta arbete är forward selection och backward elimination. Där båda metoder är ofta återkommande för modellbygge inom regressionsanalysen. (Agresti, 2013, s. 209) Procedurerna baseras till stor del på statistisk signifikans. Viktigt att notera att statistisk signifikans inte bör vara de

enda kriterium man utgår ifrån huruvida man ska exkludera eller inkludera en kovariat. Kort sagt är det är det av intresse att inkludera kovariater som spelar en central roll för utfallet även om de inte är statistiskt signifikanta. (Ibid s.210)

3.3.1 Forward Selection

I denna metod undersöker man kovariaterna en och en. Där man sedan i olika sekvenser adderar den kovariat med lägst p-värde och fortsätter på samma sätt tills man hittat den modell som passar de kriterium man har. Processen fortgår så länge vi har en signifikant modell. I varje steg då vi tillför en kovariat tittar man även på kovariater som man tillfört i tidigare steg för att se om de fortfarande är signifikanta då vi tillför variabler. (Agresti, 2013, s. 210)

3.3.2 Backward elimination

Metoden går ut på att man börjar med en komplex modell med alla kovariater av intresse där man sedan plockar bort kovariater i olika sekvenser.

I varje steg plockar man bort den kovariat som har minst effekt på modellen. Med andra ord så plockar vi bort den kovariat med högst p-värde och anpassar sedan en ny modell utan denna. Processen upprepas ända tills vi bara har signifikanta variabler kvar. (Agresti, 2013, s.210) Då man genomför denna metod där de viktigt att tittat på huruvida variablerna är korrelerade med varandra. När man bara har signifikanta variabler kvar sätter man tillbaka kovariaten man plocka bort först och fortsätter sedan i tur och ordning med de andra kovariater som exkluderats från modellen. Anledningen är att kovariaten man plocka bort först kan ha samvarierat med kovariaten man plocka bort efter. Återinförs den utan den andra kovariaten kan den absolut visa sig vara signifikant. (Ibid, s.208)

3.4 Goodness of fit

'Goodness of fit' av en statistisk modell beskriver hur väl den statistiska modellen i fråga passar den uppsättning observationer vi arbetar med.

3.4.1 Akaike informations kriterium (AIC)

Akaike informations kriterium (AIC) bedömer en modell på hur ”nära” deras anpassade värdena är de sanna medelvärdena, i termer av ett visst förväntat värde.

En optimal modell har så ”nära” anpassning som möjligt till de sanna värdena. Enligt definition vill man välja en modell som minimeras av:

AIC=-2(maximerade likelihood-antalet parametrar i modellen)

Har man exempelvis många potentiella kovariater kan man använda AIC som mått för variabelvalet till modellen. Man söker då efter den modell med minst AIC. Viktigt att notera att modeller som är väldigt snarlik den modell med minst AIC kan vara av intresse. (Agresti, 2013, s. 212)

3.4.2 Hosmer-Lemeshow

Hosmer-Lemeshow test är ett ’goodness of fit’ test anpassade för logistiska regressionsanalyser. Denna strategi lämpar sig bäst för modeller med relativt få kovariater. Då antalet kovariater ökar, minskar strategins effektivitet. Med hjälp av detta test undersöker man anpassningen hos en logistisk regressionsmodell.

Testet går ut på att man grupperar de skattade sannolikheterna $\left(\pi(\hat{x}_1) \dots \pi(\hat{x}_n)\right)$ för att ett visst utfall ska inträffa där $\pi(x) = \frac{exp(\alpha+\beta x)}{1+exp(\alpha+\beta x)}$. Man rangordnar de minsta sannolikheterna med positivt utfall och sedan i fallande ordning. (i denna uppsatts skulle positivt utfall vara att studenten klarade sig samma termin dvs Y=1)
Man delar upp de på G grupper där G=10 är ”standardgruppering” vilket vi använder oss av i denna uppsats. Den första gruppen innehåller de minsta värdena av $\pi(\hat{x}_n)$ och den sista gruppen innefattar de med störst värde. I dessa G=10 grupper kommer man att kunna observera hur många Y=1 respektive Y=0 som finns. Med andra ord kan vi observera hur många studenter som faktiskt klarade sig samma termin respektive vid ett senare tillfälle. De förväntade antalet av $\widehat{Y} = 0$ och $\widehat{Y} = 1$ jämförs med hur många Y=0 respektive Y=1 man observerat i ett visst intervall. De förväntade värdena jämförs med de observerade med hjälp av ett χ^2-test. Med χ^2-testet testar man huruvida det är stora skillnader mellan grupperna ”förväntade” och ”observerade”. Nedan i tabell 2 har vi skapat en tabell för att tydliggöra vad som menas med förväntade och observerade. Data är baserat på den logistiska regressions modell vi kommit att kalla modell 1.

Vi nämnde ovan att man med hjälp av detta test undersöker anpassningen hos en logistisk regressionsmodell. Ger Hosmer Lemeshow testet oss ett p-värde över 0.05

indikerar de om att modellen vi anpassat passar data väl medan om p-värdet understiger 0.05 tyder på bristande anpassning för modellen i fråga. Värdet 0.05 är inte skrivit i sten utan snarare ett konventionellt tröskelvärde för vart p-värdet ska vara statiskt signifikant eller inte. I detta arbetet använder vi 0.05. När man säger att modellen passar data väl innebär de att skillnaden mellan förväntade och observerade antal är liten. Har modellen istället dålig passning innebär de att skillnaderna mellan förväntade och observerade är genomgående stora. (W.Hosmer & Lemeshow, 2000, s.147-156)

I tabell 2 nedan ser vi de skattade sannolikheterna för att klara av kursen samma termin. I mitten kolumnen 'förväntade' kan vi se hur många studenter man förväntar sig ska klara sig samma termin respektive vid ett senare tillfälle. I kolumnen längst till höger 'observerade' är antalet studenter för vardera utfall vi observerat i intervallen för de skattade sannolikheterna.

Tabell 2, Hosmer-Lemeshow test baserat på modell 1

Skattade sannolikheter	**Förväntade Frekvenser**		**Observerade Frekvenser**	
	$\widehat{Y} = 0$	$\widehat{Y} = 1$	$Y = 0$	$Y = 1$
[0.166, 0.501]	35.72	25.29	32	29
(0.501, 0.573]	27.06	32.94	27	33
(0.573, 0.636]	23.43	36.57	23	37
(0.636, 0.676]	20.57	39.43	30	30
(0.676, 0.705]	18.82	42.18	19	42
(0.705, 0.732]	16.36	41.64	17	41
(0.732, 0.758]	16.17	47.83	18	46
(0.758, 0.790]	16.83	59.17	10	66
(0.790, 0.810]	8.60	35.39	6	38
(0.810, 0.943]	7.45	47.55	9	46

3.4.3 ROC & AUC

ROC- Receiver operating characteristic

Genom att betrakta alla möjliga värden för tröskelvärdet ξ mellan 0 och 1, kan ROC-kurvan konstrueras som en plott av sensitivitet (TPR-true positive rate) och 1-specificiteten (FPR-false positive rate). Där TPR är proportionen av Y=1 som blir korrekt klassificerade och FPR proportinen av Y=0 som blivit felaktigt klassificerade. Y=1 och Y=0 är den binära responsen för huruvida studenter klarar sig samma termin respektive vid ett senare tillfälle under perioden. När TPR och FPR plottas mot varandra erhålls ROC-kurvan.

ROC-kurvor är ett jämförelsemått som innefattar en grafisk illustration av modellen i fråga. Man kan då summera modellens prediktiva styrka för alla möjliga tröskelvärden mellan 0 och 1.

En student klassificeras som positiv om den hamnar över ett visst tröskelvärdet och negativ om den hamnar under. En ”sann positiv” är en student som klarade sig samma termin medan en ”sann negativ” är en student som inte klara sig samma termin utan vid ett senare tillfälle. Beroende på var tröskelvärdet är så kan en ”sann negativ” klassificieras som positiv om de hamnar över tröskelvärdet och är då en ”falsk positiv”.

TPR och FPR definieras som:

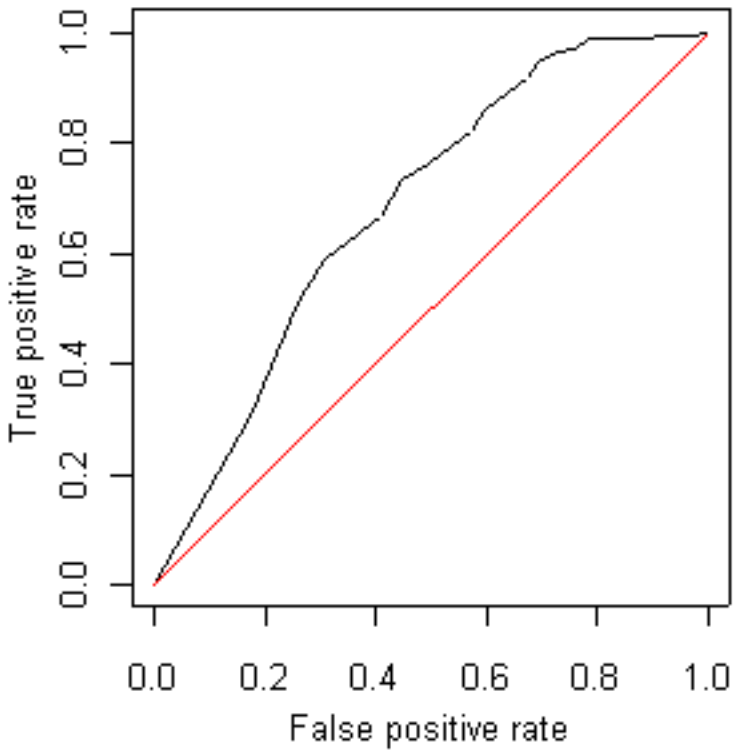
$$TPR = \frac{Antalet\ positiva\ klassade\ som\ positiva}{Antalet\ sanna\ positiva}$$

$$FPR = \frac{Antalet\ negativa\ klassade\ som\ positiva}{Antalet\ sanna\ negativa}$$

Som nämndes tidigare är TPR sensitiviteten och FPR 1-specificiteten vilket vi kan se i figur 4 nedan. Figur 4 är en illustration över hur en ROC-kurva kan sen ut.

(Agresti, 2013, s.224) (Zon, Liu, Bandos, Ohno-Machado, & Rockette, 2012, s. 6-11)

Figur 4, exempel på ROC-kurva
Källa: Se figurförteckning efter referenslista

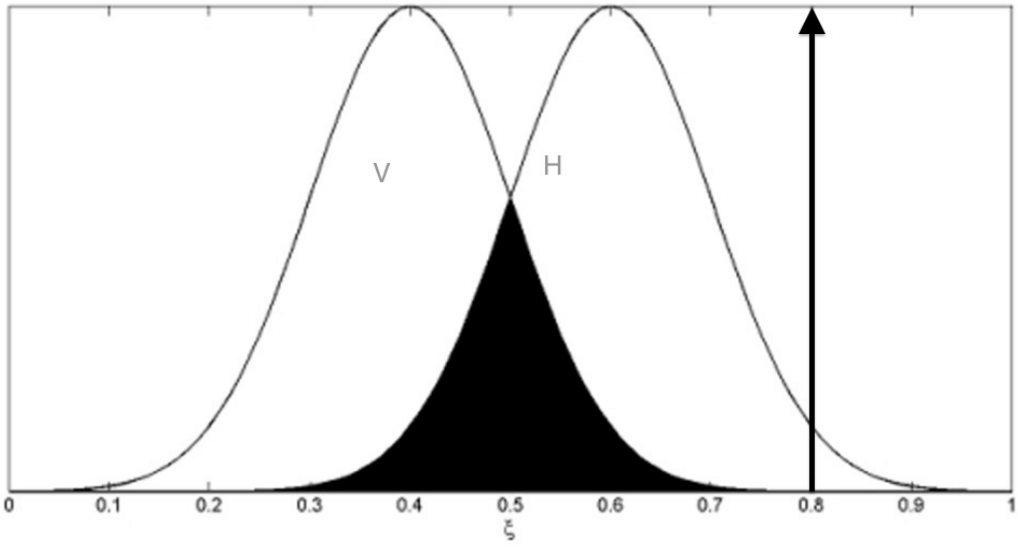


AUC, arean under kurvan är ett mått på hur bra en modell är på att diskriminera mellan två grupper. Desto större area, desto bättre. Maximum för arean är 1.0 och. ett värde på 0.5 innebär att modellen är slumpmässig dvs inte bättre än att singla slant. Just av den anledningen brukar linjen som delar figur 4 på diagonalen kallas slump/chanslinjen. En tumregel brukar vara att en ROC-kurva med area: Större än 0.9 är perfekt, 0.8-0.9 utmärkt, 0.7-0.8 acceptabel. (W.Hosmer & Lemeshow, 2000, s.160-164)

Vi använder ett exempel för att illustrera hur ROC-kurvor är konstruerade. Vi utgår från ett exempel baserat på 600 studenter dvs 600 observationer. I figur 5 nedan kan vi se H (Höger), som är antalet studenter som klarade sig samma termin respektive V (Vänster) som är antalet studenter som inte klarade sig samma termin, pilen svarar mot ett tröskelvärde ξ på 0.8. Tröskelvärdet skulle kunna placeras var som helst på x-axeln. Men i detta exempel har vi den på 0.8.

Figur 5 har inget att göra med vårt datamaterial utan är enbart till för att illustrera hur det skulle kunna se ut.

Figur 5, två normalfördelningskurvor
Källa: Se figurförteckning efter referenslista



H- de studenter som klara sig samma termin, 300 stycken.

V- De studenter som inte klara sig samma termin utan vid ett senare tillfälle, 300 stycken.

En student som hamnar till höger om tröskelvärdet klassificeras som positiv och till vänster om tröskelvärdet som negativ, oavsett om man klarat sig samma termin eller inte. Baserat på tröskelvärdet tänker vi oss att de är 50 stycken studenter som ligger till höger om själva tröskelvärdet. De 50 studenterna är 50 studenter klassade som positiva och som faktiskt har klarat sig samma termin. Inga sanna negativa har klassificerats som positiva.

För att beräkna sensitiviteten(TPR) och specificiteten(FPR) använder vi definitionen:

$$TPR = \frac{Antalet\ positiva\ klassade\ som\ positiva}{Antalet\ sanna\ positiva} = \frac{50}{300} = 0.17$$

$$FPR = \frac{Antalet\ negativa\ klassade\ som\ positiva}{Antalet\ sanna\ negativa} = \frac{0}{300} = 0$$

Punkten som plottas på ROC-kurvan blir då (0, 0.17)=(x, y). ROC-kurvan erhålls genom att räkna FPR och TPR för alla möjliga tröskelvärde.

4. Resultat

I denna del kommer vi att tillämpa metoderna ovan på datormaterialet.

Vi kommer presentera de resultat vi kommit fram till för huruvida en student klarar sig samma termin eller blir eftersläpandes med kursen och klarar den vid ett senare tillfälle perioden 2007-2014.

4.1 Modellkonstruktion

Man kunde redan innan ana vilka variabler som skulle vara intressanta att ha med i analysen för huruvida man klarade sig samma termin eller på en omtentamen vid ett senare tillfälle under angivna perioden.

För att bygga den logistiska regressionen användes till en början Backward elimination och forward selection. Detta för att ge en indikation på vilka variabler som har stor betydelse för modellbygget. Dessa metoder är baserade på p-värde.

Vi använde även den inbyggda funktionen i R som genomför backward elimination och forward selection baserade på AIC kriterium. Då man använder kriterium baserde på AIC vill man minimera AIC, dvs hitta den modell med lägst AIC. (stepAIC,i paketet MASS, 2002)

4.2 Modell 1

I denna modell har vi att tittat på studenter som var registrerade på ett kandidatprogram under perioden 2007-2014. Som vi nämnde i ’beskrivning av data’ har vi begränsat oss till enbart studenter som någon gång under angivna perioden registrerat en godkänt resultat på kursen. Vi uteslöt även datalogistudenterna då de läser kursen uppdelat på år 1 och 2 samt de fristående kursarna då de inte finns information i ladok om huruvida studenten läste på hel eller halvfart.

Vi beslöt oss sedan för att bygga den logistiska regressionen baserat på Backward elimination och forward selection baserat på p-värde. Anledningen till de beslutet var att modellen vi fick då vi minimerade AIC hade ett AIC på 715.63. När vi istället baserade vårt test på p-värde fick vi en modell med ett AIC på 715.97. De två modellerna hade således snarlika AIC värde. Backward elimination och forward selection baserade på AIC ville inkludera kovariaterna ålder, program, kalenderår samt förberedande kurs medan samma metoder baserade på p-värde inkluderade samma kovariater exklusive förberedande kursen. Variabeln förberedande kurs är

som vi kan se i tabell 3 inte signifikant. Vi valde således att utgå från backward elimination och forward selection baserade på p-värde.

De 599 Studenterna analyserades med kovariaterna program, ålder vid registrering, kön, kalenderår och huruvida man gått förberedande kurs. Terminsstarten exkluderades innan vi genomförde en logistiska regressionen då de inte är intressant om studenterna började kursen under höst eller vårtermin då man läser ett kandidatprogram eftersom terminsstarten bestäms av kursplanen. I tabell 3 nedan ser vi resultatet av den logistiska regressionen med alla intressanta kovariater inkluderade.

Tabell 3, Resultat från logistisk regression med alla kovariater av intresse

			Konfidensintervall för β	
	β	**P-värde**	**2.5%**	**97.5%**
Program				
Lärare	1.34	0.008	0.41	2.42
Astronomi	-0.17	0.64	-0.87	0.56
Biofysik	0.91	0.42	-0.99	3.89
Fysik	0.38	0.19	-0.19	0.97
Meteorologi	-0.08	0.84	-0.83	0.70
Matematik och filosofi	-0.26	0.59	-1.19	0.72
Sjukhusfysiker	0.59	0.14	-0.17	1.43
Matematik och ekonomi	-0.49	0.07	-1.03	0.04
Biomatematik	-1.08	0.01	-1.93	-0.25
Matematik	0	0	0	0
Ålder	-0.07	0.0005	-0.10	-0.03
Kön				
Kvinna	0	0	0	0
Man	-0.06	**0.74**	-0.43	0.31
Förberedande kurs				
Ja	-0.42	0.16	-0.99	0.17
Nej	0	0	0	0
Kalenderår	-0.09	0.06	-0.19	0.002
AIC:	723.44			

I första steget eliminerades kön som hade ett högt p-värde (0.74). Program som inte var signifikant skilda från matematikprogrammet slogs ihop med basklassen. Basklassen var således matematikprogrammet. I tabell 4 nedan kan vi se modellen då vi exkluderat kön samt slagit samman programmen som inte var signifikant skilda från matematikprogrammet.

Tabell 4, resultat av logistisk regression då kön eliminerats

			Konfidensintervall för β	
	β	**P-värde**	**2.5%**	**97.5%**
Program				
Lärare	1.22	0.01	0.35	2.27
Biomatematik	-1.21	0.002	-2.01	-0.43
Matematik och ekonomi	-0.60	0.01	-1.07	-0.14
Matematik	0	0	0	0
Ålder	-0.07	0.0004	-0.10	-0.03
Förberedande kurs				
Ja	-0.45	**0.12**	-1.02	0.13
Nej	0	0	0	0
Kalenderår	-0.11	0.02	-0.21	0.01
AIC:	715.63			

I detta steget eliminerade vi även förberedande kursen, med ett p-värde på 0.1228. I tabell 5 nedan kan vi se hur modellen ser ut då vi eliminerat alla kovariater med höga p-värde. Det är denna modell vi kommer utgå från då vi tittar närmre på resultaten. Som vi ser så finns det bara signifikanta variabler kvar. Låt oss kalla denna modell för modell 1.

Tabell 5, resultat av logistisk regression då förberedande kurs eliminerats

			Konfidensintervall för β	
	β	**P-värde**	**2.5%**	**97.5%**
Program				
Lärare	1.23	0.01	0.36	2.28
Biomatematik	-1.16	0.004	-1.95	-0.38
Matematik och ekonomi	-0.59	0.01	-1.05	-0.12
Mateatik	0	0	0	0

Ålder	-0.07	0.0003	-0.11	-0.03
Kalenderår	-0.12	0.01	-0.22	-0.03
AIC:	715.97			

I tabell 6 ser vi en tabell med oddskvoterna för modell 1.

Tabell 6, Oddskvoter baserat på resultatet i tabell 5. (Modell 1)

		Konfidensintervall för oddskvoten	
	Oddskvot	**2.5%**	**97.5%**
Program			
Lärare	3.43	1.43	9.74
Biomatematik	0.31	0.14	0.68
Matematik och Ekonomi	0.56	0.35	0.89
Ålder	0.93	0.90	0.97
Kalenderår	0.89	0.81	0.97

Med den logistiska regressionen identifierades en intressant modell som innefattade kovariaterna program, ålder samt kalenderår, vilket vi kan se i tabell 5. Vi valde matematikprogrammet som basklass. I den första logistiska regressionen som vi ser i tabell 3 kan man se hur de andra kandidatprogrammen presterar i förhållande till matematikprogrammet. Programmen som presterade bättre baserat på att klara sig samma termin var lärare, biofysik, fysik samt sjukhusfysiker. Kandidatprogrammen som i förhållandevis presterade sämre var astronomi, biomatematik, meteorologi, matematik/filosofi samt matematik/ekonomi. Program som inte var signifikant skilda från matematikprogrammet slogs samman med basklassen. Som ses i tabell 3 ovan var de lärare-, biomatematik- samt matematik och ekonomiprogrammet som var signifikanta.

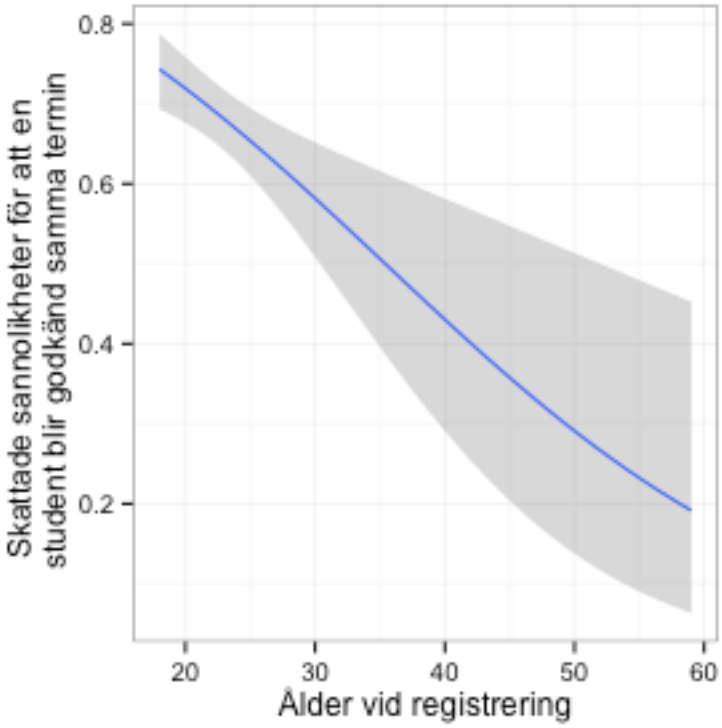
Baserat på oddskvoter i tabell 6 ovan kan vi säga att för en enhets ökning dvs. att man ökar ålder med ett år är oddskvoten 0.9347 och oddset att studenten klarar sig samma termin är då 6.5 % lägre.

Nedan i figur 7 ser vi en plott tillsammans med ett 95 % konfidensintervall för de skattade sannolikheterna att en student blir godkänd samma termin baserat på ålder

vid registrering. På y-axeln ser vi skattade sannolikheterna för att en student blir godkänd samma termin och på x-axeln ser vi studentens ålder vid registrering. Vad vi ser i figur 7 är att studenter som påbörjar studierna direkt efter gymnasiet har en större andel godkända samma termin. Bland de studenterna som undersöktes var ålders intervallet 18-59 år.

För figur 7 och 8 nedan användes programpaket i R. (ggplot i paketet ggplot2, 2009) Konfidensintervallet skapas genom att ggplot2 använder prediktionsfunktionen för att ta fram standardfelet för varje prediktion. Våra skattade log odds för huruvida man klarar sig samma termin är approximativt normalfördelade så felmarginalen för konfidensintervallet beräknas som $\pm$ 1.96*standardfelet. Kovariaterna hålls till sina basklasser. Konstruktionen för konfidensintervallet gäller för både figur 7 och 8 nedan.

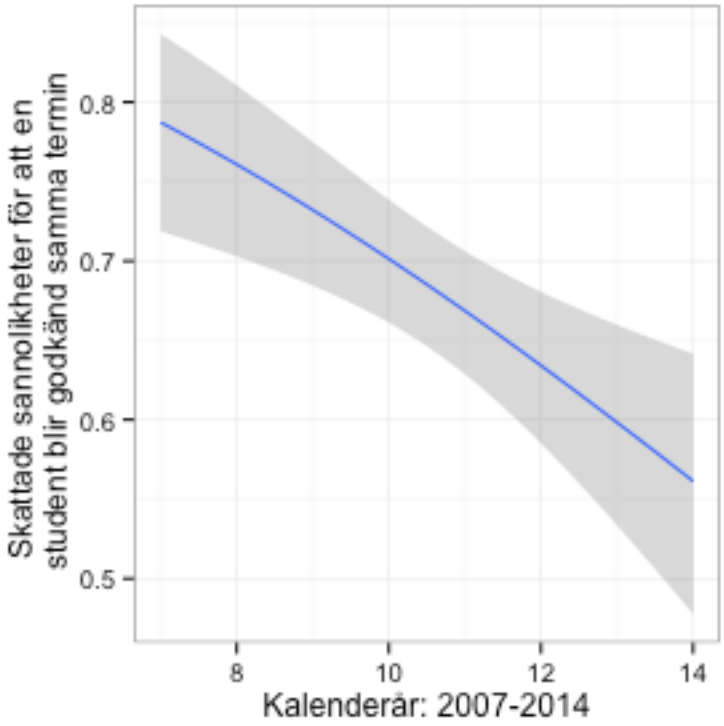
Figur 7, Skattade sannolikheter för att bli godkänd samma termin baserat på ålder vid registrering



För kalenderåret hade vi en oddskvot på 0.8861 och går vi ett kalenderår fram exempelvis från 2007 till 2008 är oddset att studenten klarar sig samma termin 11.4% lägre. I figur 8 nedan har vi en en plott tillsammans med ett 95% konfidensintervall för de skattade sannolikheterna att en student klarar sig samma termin baserat på vilket kalenderår man var registrerad. Kovariater hålls till sina

basklasser. På y-axeln ser vi skattade sannolikheterna för att en student blir godkänd samma termin och på x-axeln ser vi kalenderåren som studenterna varit registrerade på. Vad vi ser är att andelen som klarade sig samma termin 2007 var strax under 80% som minskat fram till 2014 där andelen som klarar sig samma termin istället ligger strax över 55 %.

Figur 8, Skattade sannolikheter för att bli godkänd samma termin baserat på vilket kalenderår studenten var registrerad



Som nämndes i 'beskrivning av data' har de som registrerades 2007 haft längre tid på sig att avsluta kursen än de som registrerades senare under perioden fram till 2014. Var det inga skillnader i prestationer mellan kalenderåren bör rimligtvis kurvan vara växande istället för avtagande. Som ses i figur 8 är andelen godkända samma termin högre längre bak i tiden, vilket tyder på att genomströmningen för studenter som klarat sig samma termin avtagit under perioden. Vilket också betyder att andelen eftersläpande studenter för varje termin ökat.

De olika programmen tolkas på likartat sätt. Oddskvoten för matematik och ekonomiprogrammet är 0.5559 vilket betyder att oddset för att en student klarar sig samma termin är ca 45 % lägre än för en student som läser på matematikprogrammet. Samma sak gäller för biomatematik studenterna, oddset för att de ska klara sig samma termin är 69% lägre än för de som läser matematik. De

går bättre för lärarstudenterna, oddset att de ska klara sig samma termin är 3.42 gånger högre än att en matematikstudent gör det.

4.3 Hosmer-Lemeshow

Vi genomförde ett Hosmer-Lemeshow med hjälp av programpaket i R. (hoslem.test i paketet ResourceSelection, 2014)

Med hosmer-lemeshow test undersöker man anpassningen hos en logistisk regressionsmodell, dvs hur väl modellen passar de data vi har. När man säger att modellen passar data väl innebär de att skillnaden mellan förväntade och observerade värden är liten. Har modellen bristande anpassning innebär de istället att skillnaden mellan oberserverade och förväntade värden är stor. Ger testet oss ett p-värde över 0.05 indikerar de om att modellen vi anpassat passar data väl. Medan om p-värdet understiger 0.05, tyder på bristande anpassning för modellen i fråga. Värdet 0.05 är inte skrivit i sten utan snarare ett konventionellt tröskelvärde för vart p-värdet ska vara statiskt signifikant eller inte. I detta arbetet använder vi 0.05. I nedanstående tabell 7 kan vi se resultatet för genomförande av Hosmer-Lemeshow testet baserat på vår modell 1.

I detta fallet har vi som ses nedan i tabell 7 ett p-värde på 0.449. Då p-värdet överstiger 0.05 har modellen god anpassning.

Tabell 7, Resultat av Hosmer-Lemeshow test

Modell	χ^2	Frihetsgrader	P-värde
Modell 1	7.84	8	0.45

4.4 ROC & AUC

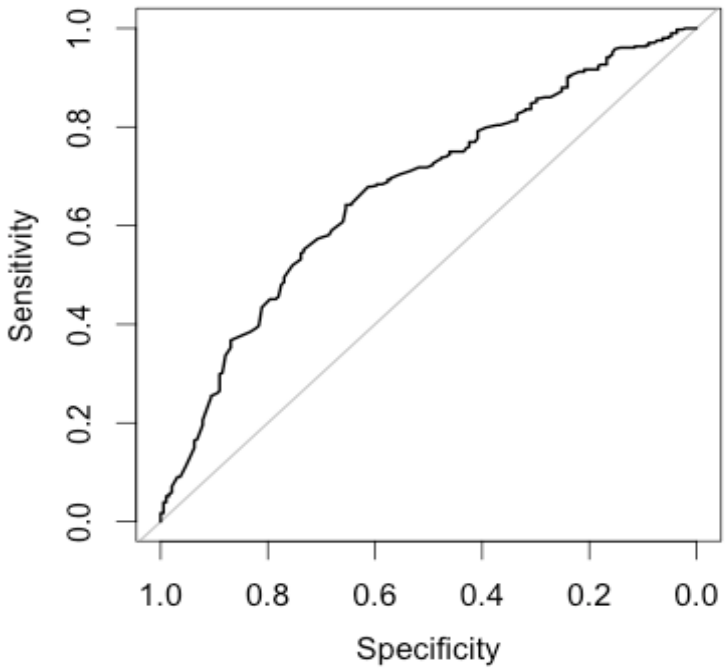
Vi använde oss av ROC för att undersöka modellens prediktiva förmåga. För att konstruera vår ROC-kurva användes ett programpaket i R. (roc i paketet pROC, 2011)

Den binära responsen är: studenten klara sig samma termin/studenten klara sig inte samma termin utan vid ett senare tillfälle. Där en positiv student är en som klarade

sig samma termin och negativ student är en som inte klara sig samma termin. I figur 9 nedan kan vi se den ROC-kurva som skapades. Sensitiviteten (TRP)plottas mot 1-specificiteten (FPR). Där sensitiviteten är en student som klara sig samma termin och blir korrekt klassificerad och specificiteten är en student som inte klarade sig samma termin blir felaktigt klassificerad. Specificiteten är de studenter som inte klarade sig men klassificerats som positiva.

Arean under kurvan mäter diskrimineringsförmågan hos modellen i fråga dvs att korrekt kunna klassificera de studenter som klarar sig samma termin och de som inte gör det.

Figur 9, ROC-kurva för modell 1



I tabell 8 kan vi se att arean under kurvan i figur 9 är 0.6765. Vilket varken är strålande eller värdelöst. Vi kan således baserat på arean under kurvan säga att de finns svårigheter i att predicera för huruvida en student klarar sig samma termin eller vid ett senare tillfälle.

Tabell 8, Resultat av arean under ROC-kurvan

Modell	Area under kurva (AUC)
1	0.68

5. Diskussion

De 599 studenter vi analyserade var godkända studenter bland 1547 studenter som var registrerade på ett kandidatprogram under höstterminen 2007 och vårterminen 2014. Som nämnts tidigare har vi begränsat oss till program studenter som läser på helfart.

Bland de 599 studenter vi analyserade var det 408 studenter som totalt klarade sig samma termin. Sannolikheten att man klarar sig samma termin ligger på 68.6 % givet att de andra studenterna klarar sig vid ett senare tillfälle. 68.6 % kan låta förvånansvärt högt men om man istället beräknar sannolikheten att en student klarar sig samma termin baserat på antalet registrerade så är sannolikheten att klara sig samma termin bara 26%. Resterade 74 % blir således eftersläpandes med ett eller flera moment kvar av kursen. Kort sagt har institutionen ett problem med eftersläpande studenter.

I denna analys har vi konstaterat att åldern spelar roll för huruvida studenter uppnår ett godkänt betyg samma termin. Där andelen yngre studenter klarar sig bättre. Yngsta studenten är 18 år i materialet och de har visat sig att desto yngre man är desto större är chansen att man får ett godkänt betyg samma termin. Det är sedan tidigare känt att det är betydligt vanligare att göra avbrott i studierna i högre ålder än yngre. Man blev registrerad på kursen och kom tillbaka x antal terminer senare och klarade kursen då. Baserat på en rapport från universitet kanslers ämbetet är de ca 50-60 % som gör avbrott i studierna om man är över 35 års ålder. Medan de i åldrarna 25-34 ligger på mellan 20-30 % samt för studenter i åldern 18-24 under 10%. (Amnéus & Gillström, 2008)

Baserat på data som finns tillgänglig för denna uppsats har vi inte kunnat statistiskt säkerställa vad som beror på att andelen godkända samma termin sjunker då man blir äldre.

Vi vet inte om de har att göra med att yngre studenter som går direkt från gymnasiet klarar sig bättre av den anledningen och att det tvärt om skulle vara så att de äldre studenterna börjat glömma och det är därför det går sämre. Kanske förklaringen rentav ligger utanför universitetet där det är tänkbart att äldre studenter har större

ansvar för eventuella barn, familj och ekonomiska situation. Som nämndes ovan är detta enbart spekulationer och inget vi har kunnat statistisk säkerställa.

Kalenderåret spelade också roll för utfallet, där det visade sig att andelen godkända samma termin sjunker med tiden. Den nedåtgående trenden måste förklaras av någon bakomliggande faktor så att de inte är kalenderåret i sig som gör att de är förre andel studenter som klarar sig samma termin. Annars hade vi om x antal år inte haft en enda student som blev godkänd samma termin.
Vad vi noterat i samband med detta är att studentkullarna ökat rejält sedan HT 07 då det registrerades 170 studenter medan de VT 14 registrerades hela 460 studenter. Vi har även noterat att de är förre studenter som registreras på våren, HT 13 hade kursen 501 studenter registrerade.
Vi undersökte denna saken närmare för att möjligtvis hitta en förklaring till den nedåtgående trenden. Vi skapade kovariaten 'antal registrerade per termin' och utförde logistisk regression. Anledningen till att vi delade upp de till antal registrerade per termin istället för antal registrerade per kalenderår har och göra med att vi enbart ha data för vårterminen 2014. När vi utförde logistiska regressionen uteslöt vi kalenderåret då de visade sig vara starkt korrelerade med varandra. Det visade sig dock inte förklara särskilt mycket.
Vad vi undersökte var om årskullar (termin för termin) med fler antal registrerade studenter presterade sämre.
Oddskvoten låg på 0.9987 vilket betyder att oddset då man tillför ytterligare en student blir oddset 0.13 % lägre att man klarar sig samma termin. Denna variabel var inte heller signifikant i modellen.
Förklaringen till den nedåtgående trenden kan således inte förklaras av att studenterna blir "mindre sedda" i de stora kullarna.

Vi kollade även om kursupplägget hade förändrats, vilket det inte hade gjorts.

Då åldern hade en negativ påverkan, desto äldre man var desto mindre troligt var de att man blev godkänd samma termin så undersöktes det om de var så att kalenderåren hade en avvikande åldersfördelning. Några sådana samband fanns inte och vi hade en medianålder på ca 21 år igenom hela tidsperioden.

Vilket program man var student vid hade också betydelse för om man var en student som klarade sig samma termin eller bidrog till eftersläpet. Vad de gäller

kandidatprogrammen så har de olika antagningskrav för att kunna bli antagen till programmet. För programmen som presterar bättre i förhållande till matematikprogrammet kan man se att de kräver även fysik A, fysik B, kemi A och kemi B som förkunskaper. Tillskillnad från matematikprogrammet där man enbart har matematik D som förkunskapskrav. Programmet som presterar sämst i förhållande till matematikprogrammet är matematik och ekonomiprogrammet. Vid en närmre titt ser vi att förkunskapskraven från 2007-2014 var enbart matematik C, vilket sedan förändrades till matematik D. Vi kan således inte se om de gått bättre för studenterna efter den förändringen då materialet bara sträcker sig fram till vårterminen 2014, förändringen gällde studenterna som började på höstterminen 2014. Detta skulle kunna vara en förklaring till varför de går avsevärt sämre för de studenter inom den tidsperioden. Detta är inget som är statistiskt säkerställt utan enbart en spekulation.

De variabler som reducerades från modellen då de visade sig vara icke signifikanta för utfallet var kön samt huruvida man gått förberedande kurs. Kan vara intressant att kommentera vad de hade för effekt trotts att vi uteslutit de från modellen. Baserat på logistiska regressionen i tabell 3 visade det sig att oddset för att en man ska klara sig samma termin är 6 % lägre än för kvinnor. Den förberedande kursen visade ett minst sagt intressant resultat där det visade sig att om man läste kursen så var oddset 37 % lägre att klara sig samma termin. Bland de 599 studenter som analyserades var det 58 studenter som hade läst den förberedande kursen. Detta baserat på logistiska regressionen i tabell 3. För nyfikenhetens skull kollade vi upp ”vilken typ” av studenter som hade läst den förberedande kursen. Där fanns inget anmärkningsvärt. Möjligen är det studenter som känner sig svaga redan innan terminsstart som väljer att gå den. Vilket sedan speglar resultatet för huruvida man klarar kursen samma termin. Ovanstående spekulationer kring vad som ligger bakom förberedande kursen är inte statistiskt säkerställt.

Referenser

Agresti, A. (2013). *Categorical Data Analysis.* United States of America: John Wiley & Sons, Inc.,Hoboken, New Jersey.

Amnéus, I., & Gillström, P. (2008). *Vilka är studenter? En undersökning av studenterna i Sverige.* Stockholm: Högskoleverket.

Ehlin Kolk, M. (2012). *Vi måste våga prata om genomströmning i högre utbildning.* Almedalen: www.saco.se.

Matematiska, Instutionens hemsida. (den 20 Mars 2015). *Matematiska Institutionen*. http://www.math.su.se/utbildning/kurser/matematik-i den 30 April 2015

Rolf, S. (2014). *Lineära Statistiska Modeller*. Stockholm: Matematiska Inst.

W. Hosmer, D., & Lemeshow, S. (2000). *Applied logistic regression.* Canada: Wiley 2000.

Zon, K. H., Liu, A., Bandos, A. I., Ohno-Machado, L., & Rockette, H. E. (2012). *Statistical Evaluation of diagnostic performance topics in ROC analys.* Boca Raton: Taylor & Francis.

Figur 4, s.13: Figuren visar en ROC-kurva. Hämtad 2015-05-09.
http://www.epa.gov/caddis/pecbo_estimating6.html

Figur 5, s. 14: Figuren visar två normalfördelningskurvor. Hämtad 2015-05-09.
http://www.biomedcentral.com/1471-2148/10/137/figure/F4?highres=y

Figur 6, s.15: Figuren visar en ROC-kurva. Punkten är egenkonstruerad.
Hämtad 2015-05-09.
http://blogs.sas.com/content/iml/2011/06/03/a-statistical-application-of-numerical-integration-the-area-under-an-roc-curve.html

R MASS package: Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

R ggplot2 package: H. Wickham. ggplot2: elegant graphics for data analysis. Springer New York, 2009.

R ResourceSelection package: 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

R pROC package: Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez and Markus Müller (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, p. 77. DOI: 10.1186/1471-2105-12-77 <http://www.biomedcentral.com/1471-2105/12/77/>

Appendix

Fördelningen för kovariaterna modell 1

Kovariater		Antal(procent)
Kön	Kvinna	300 (50)
	Man	299 (50)
Terminsstart	Höst	469 (78)
	Vår	130 (22)
Kalenderår	HT 07- VT 14	599 (100)
Ålder	18-59	599 (100)
Program	Lärare	41 (7)
	Astronomi	46 (8)
	Biofysik	6 (1)
	Biomatematik och beräkningsbiologi	29 (5)
	Fysik	109 (18)
	Matematik	151 (25)
	Meteorologi	41 (7)
	Matematik och filosofi	22 (4)
	Sjukhusfysiker	55(9)
	Matematik och ekonomi	99 (16)
Totalt		**599**

Ålders fördelning per kalenderår

	Ålder vid terminsstart		
Kalenderår	**Min**	**Max**	**Median**
2007	18	38	20.5
2008	18	47	20
2009	18	49	21
2010	18	59	21
2011	18	59	22
2012	19	34	21
2013	19	46	21
2014	19	40	24.5

Fördelning för kovariaterna för de som läst förberedande kursen

Kovariater		**Antal(procent)**
Kön	Kvinna	20 (35)
	Man	38 (65)
Kalenderår	HT 07- VT 14	58 (100)
Ålder	19-39	58 (100)
Program	Lärare	4 (7)
	Astronomi	5 (9)
	Biofysik	1 (2)
	Fysik	9 (15)
	Matematik	22 (38)
	Meteorologi	2 (3)
	Matematik och filosofi	3 (5)
	Sjukhusfysiker	2 (3)
	Matematik och ekonomi	10 (18)
Totalt		**58**

Resultat från logistisk regression då vi bytte kalenderår till antal registrerade per termin

			Konfidensintervall för β	
	P-värde	β	**2.5%**	**97.5%**
Program				
Lärare	0.0106	1.2944	0.3659	2.3766
Astronomi	0.6574	-0.1608	-0.8631	0.5653
Biofysik	0.4273	0.8948	-1.0030	3.8824
Fysik	0.1588	0.4153	-1.1558	1.0030
Meteorologi	0.8806	-0.0578	-0.7975	0.7192
Matematik och filosofi	0.6010	-0.2518	-1.1862	0.7219
Sjukhusfysiker	0.1052	0.6530	-0.1063	1.4869
Matematik och ekonomi	0.0300	-0.5949	-1.1355	-0.0586
Biomatematik	0.0122	-1.0650	-1.9125	-0.2355
Ålder	0.0003	-0.0689	-0.1074	-0.0323
Kön				
Kvinna				
Man	0.7810	-0.0518	-0.4180	0.3138
Förberedande kurs				
Ja	0.1235	-0.4570	-1.0362	0.1314
Nej				
Antal registrerade	0.2867	-0.0009	-0.0026	0.0008
AIC:	725.99			