Lunds Universitet
Statistiska institutionen

Examensgraden vid Yrkeshögskolan i Sverige 2016 - 2020

- En multipel logistisk regressionsanalys

Författare: Olivia Peetre Malthe
Handledare: Peter Gustafsson och Johan Löfgren
Examinator: Björn Holmquist
STAH11: Kandidatuppsats, 15 hp
Seminariedatum: 2022-01-13

Abstract

This study aims to examine the factors that determine the degree of examination in higher vocational education in Sweden. Today, only around 70 percent of the students who have completed their study round receive a degree. A multiple logistic regression was applied to investigate the factors that affect the degree of examination. The method helped answer the following questions: Which factors are the most important, which ones can we ignore, how do they interact with each other, and how confident are we of these factors.

According to the investigation, there are eight vital connections with the degree of examination. The following factors are the most important characteristics to determine whether a person will complete their education or not: The field of study, gender, the form of education (distance or classroom), grades, domestic or foreign-born, the number of people applying for the spot, if they have had any occupation before the study round begins and their level of education.

Keywords: higher vocational education, degree of examination, multiple logistic regression

Förord

Jag vill tacka min handledare, Johan Löfgren, vid Statistiska centralbyrån för all hjälp med programmering och synpunkter som bidragit till värdefulla insikter. Jag vill även tacka Evalena Andersson och Paula Kossack för all hjälp med datamaterialet och för värdefull information om yrkeshögskolan.

Innehållsförteckning

1 Inledning

Den här rapporten handlar om examensgraden vid Yrkeshögskolan i Sverige. Mer specifikt undersöks vilka faktorer som samvarierar med andelen individer som examineras från yrkeshögskolan. Arbetet sker i samarbete med Statistiska centralbyrån, den statliga myndighet som har i uppdrag att samla in och presentera statistik.

1.1 Bakgrund

I Sverige erbjuds Yrkeshögskoleutbildningar inom ett flertal branscher utsträckt över hela landet. Utbildningen är eftergymnasial och kombinerar teori med praktik som utförs i nära kontakt med näringslivet. Utbildningarna syftar till att tillgodose arbetsmarknaden med den arbetskompetens som saknas. Således är utbildningarna generellt inriktade på specifika yrkesroller och dess innehåll är en reflektion av den kunskap som saknas på arbetsmarknaden. Privata företag, kommuner, landsting och universitet samt högskolor bedriver verksamheterna för yrkeshögskolan. Dessa kvalitetsgranskas av Myndigheten för yrkeshögskolan och är under statlig kontroll (Myndigheten för yrkeshögskolan, 2021a). Det finns 15 utbildningsområden och de är följande:

- Data/IT
- Ekonomi, administration och försäljning
- Friskvård och kroppsvård
- Hotell, restaurang och turism
- Hälso- och sjukvård samt socialt arbete
- Journalistik och information
- Juridik
- Kultur, media och design
- Lantbruk, djurvård, trädgård, skog och fiske
- Pedagogik och undervisning
- Samhällsbyggnad och byggteknik
- Säkerhetstjänster
- Teknik och tillverkning
- Transporttjänster
- Övrigt

Myndigheten för yrkeshögskolan presenterade år 2021 en årsrapport om examensgraden vid yrkeshögskolan. Där redogörs att examensgraden sedan år 2009 haft en uppåtgående trend. Andelen examinerade har ökat från 63 procent till dagens nivåer som ligger runt 70 procent. Ökningen gäller totalt samt för kvinnor och män var för sig. Men mellan åren 2016 och 2019 kunde en nedgång i examensgraden påvisas, mer specifikt skedde det en minskning på ca 1 procent per år (Myndigheten för yrkeshögskolan, 2021a). År 2020 ökade dock den totala examensgraden marginellt jämfört med 2019 (Statistiska centralbyrån, 2021). Generaldirektören för yrkeshögskolan skriver i ett inlägg att konsekvenserna av färre examinerade blir att arbetsmarknaden inte tillgodoses med den kompetens som efterfrågas. Dessutom minskar kostnadseffektiviteten när utfallet blir mindre kompetens förvärvad för insatta resurser (Myndigheten för yrkeshögskolan, 2021b).

Tabell 1: Antalet samt andelen individer som tagit examen respektive ej tagit examen år 2016-2020.

Examen (Y)	**antal**	**andel**
examen	71 405	72,4 %
ej examen	27 185	27,6 %

Tabell 1 visar att cirka 72 procent av de individer vars utbildningsomgång avslutats mellan 2016-2020 har tagit ut en examen. Sålunda har cirka 28 procent inte fått en examen efter avslutad utbildningsomgång.

Tabell 2: Examensgraden fördelat på kön respektive utbildningsområde 2016-2020.

	Examensgraden	
Utbildningsområde	**Män**	**Kvinnor**
Data/IT	57 %	63 %
Ekonomi, administration och försäljning	73 %	82 %
Friskvård och kroppsvård	55 %	80 %
Hotell, restaurang och turism	62 %	75 %
Hälso- och sjukvård samt socialt arbete	69 %	79 %
Journalistik och information	67 %	73 %
Juridik	51 %	74 %
Kultur, media och design	80 %	84 %
Lantbruk, djurvård, trädgård, skog och fiske	68 %	76 %
Pedagogik och undervisning	81 %	81 %
Samhällsbyggnad och byggteknik	67 %	70 %
Säkerhetstjänster	76 %	78 %
Teknik och tillverkning	61 %	68 %
Transporttjänster	80 %	79 %
Övrigt	63 %	69 %

Tabell 2 visar examensgraden år 2016-2020 inom respektive utbildningsområde fördelat på kvinnor och män. Tabellen åskådliggör att inom samtliga utbildningsområden förutom pedagogik och undervisning samt transporttjänster har kvinnor en högre examensgrad än män. Till exempel ser man att inom data/IT har 63 procent av kvinnorna tagit examen medan motsvarande siffra för män är 57 procent. Det kan även konstateras att det utbildningsområde med lägst examensgrad för kvinnor var data/IT och för män var det juridik. Kultur, media och design var det utbildningsområde med högst examensgrad för både kvinnor och män.

1.2 Syfte

Vilka faktorer samvarierar med examensgraden vid Yrkeshögskolan?

För att besvara frågeställningen kommer ett dataset bestående av individer vars utbildningsomgång avslutats mellan åren 2016-2020 att analyseras. Metoden som används för analysen är multipel logistisk regressionsanalys. Den används i syfte att finna vilka faktorer som påverkar den beroende variabeln examensgraden. Vidare används metoden för att besvara frågorna: Vilka faktorer är de mest väsentliga, vilka kan vi bortse från, hur samspelar dessa med varandra, samt hur säker är vi på dessa faktorer. Sålunda är ändamålet att fastställa vilka samband som kan föreligga mellan examensgraden och de förklarande variablerna.

2 Data

2.1 Datamaterial

Datamaterialet som rapporten grundas på kommer från Statistiska centralbyrån. Den är insamlad från en naturligt förekommande grupp, nämligen antagna individer vid yrkeshögskolan. Samtliga individer som har studerat vid yrkeshögskolan är inkluderade, det är därför heltäckande och utgör en totalundersökning. Datasetet består av 98 592 individer som påbörjat en utbildning vid en yrkeshögskola vars utbildningsomgång avslutats mellan åren 2016-2020.

Även om en individ har studerat vid yrkeshögskolan tidigare tas denne med som en ny individ var gång de påbörjar en ny utbildning. Därmed kan en och samma individ förekomma mer än en gång i datamaterialet, men då på en ny rad. Datamaterialet kommer att bearbetas i programvaran SAS 9.4 Swedish version.

2.2 Variabler

I datamaterialet ingår 52 variabler, de variabler som bedöms vara mest relevanta för utfallet examen kommer att presenteras nedan. Variablerna är både kategoriska och kontinuerliga.

Den beroende variabeln *examensgraden* är en dikotom variabel eftersom händelsen, en individ tar ut sin examen, är en händelse som antingen inträffar eller inte.

$y_i = 1$ om individen tar ut en examen
$y_i = 0$ om individen ej tar ut en examen

Sålunda är *Y* en binär variabel med två nivåer där en individ bara kan ingå i en av nivåerna. Med examensgraden avses andelen examinerade i procent av antagna individer vid yrkeshögskolan som studerat. De utbildningar som ej ger examen är inte inkluderade i variabeln. Antagna individer som studerat kurser och kurspaket är även de exkluderade.

2.2.1 De oberoende variablerna

I det här avsnittet presenteras de oberoende variablerna.

År - året som utbildningsomgången avslutats. Om en individ tar ut en examen efter att utbildningsomgången avslutats räknas den ändå in i det året som utbildningsomgången officiellt avslutas.

Ålder - deltagarens ålder i december det år mätningen genomförts.

Kön - individens juridiska kön, kodat som 1 för man och 2 för kvinna.

Studieform - bunden eller distans.

Bunden innebär att större delen av utbildningen utförs på plats. Distans innebär att större delen av utbildningen sker på annan plats än skolans anläggningar.

Inrikes_utrikes_yh - inrikes- eller utrikes född

Utbildningsområde - vilket ämne individen studerar. Det finns 15 olika utbildningsområden och dessa finns listade i avsnitt 1.1.

Jmftal_grupperad (betyg) - gymnasiebetyg omräknade i poäng, dessa är grupperade i fyra klasser: (1) *00,0-09,99*, (2) *10,0-12,49*, (3) *12,5-14,99* , samt (4) *17,5-20,00*.

UtlSvBakg_YH -

11: Utrikes född

12: Inrikes född med två utrikes födda föräldrar

21: Inrikes född men med en inrikes- och en utrikesfödd förälder

22: Inrikes född med två inrikes födda föräldrar

FodelseLand_Varldsdel_Flergen - världsdel för födelseland

Gymnasiebakgrund_klartext - om individen har avklarat ett yrkes- eller studieförberedande gymnasieprogram.

Sokande_plats - antalet sökande per plats

Behor_sök_plats_grupperad - behöriga sökande per plats indelat i 8 grupper:
(1) *00,0-01,99*, (2) *02,0-02,99*, (3) *03,0-03,99*, (4) *04,0-04,99*, (5) *05,0-05,99*, (6) *06,0-06,99*, (7) *07,0-07,99* samt (8) *8+*.

UtbildningsLangdPoangYH_Utb - utbildningens längd i yh-poäng

Startar_utb - året som utbildningsomgången startar

Avslutsar_utb - året som utbildningsomgången avslutas

StartArManad_Utb - startår samt startmånad för utbildningsomgången

AvslutsArManad_Utb - avslutsår samt avslutsmånad för utbildningsomgången

Utbnivå_sunniva_old (Utbildningsnivå) - individens utbildningsnivå före start enligt svensk utbildningsnomenklatur som har en nivå och inriktningsklassifikation (sun).

Föräldrars_utbnivå - föräldrars utbildningsnivå enligt sun.

Syss_fore (Sysselsättning_före) - visar om individen hade en sysselsättningen före påbörjad utbildning.

Syss_under (Sysselsättning_under) - visar sysselsättningen under tiden utbildningen pågår.

2.3 Partiellt bortfall

I variabeln jmftal_grupperad(betyg) finns det en stor andel uppgifter som saknas. Det antas bero på att en del individer tidigare har bott utomlands och att de därför inte har dokumenterade betyg i Sverige. Variabeln Behor_sök_plats_grupperad saknar också en stor andel uppgifter. Det beror på att uppgifterna för sökande och behöriga år 2014 inte fanns tillgängliga i den versionen av datamaterialet som författaren tagit del av.
Trots den stora andelen saknade värden behålls variablerna då de antas ha en potentiell påverkan på examensgraden. De individer som saknar uppgifter blir då indelade i en egen grupp kallad "uppgift saknas".

Bortfallen kan ha en signifikant effekt på vilken modell som slutligen väljs från datamaterialet. Problem som kan uppstå vid bortfall är följande:

1. Testets styrka minskar, det vill säga sannolikheten att förkasta en falsk nollhypotes minskar. Vidare ökar då risken för ett typ-II fel.
2. Vid skattning av parametrarna kan bias uppstå.
3. Modellen kan ge en felaktig bild av verkligheten.
4. Vid analys av modellen kan det bli mer komplicerat att dra korrekta slutsatser eftersom hänsyn måste tas till gruppen "uppgift saknas".

Samtliga punkter måste beaktas vid analysen eftersom de uppgifter som saknas kan påverka vilken modell som bedöms som mest lämplig (Kang, 2013).

3 Metoder

3.1 Multipel logistisk regressionsanalys

Syftet med arbetet är att finna ett samband mellan den beroende variabeln *y* och förklaringsvariablerna ($x_1,...,x_k$). Då den beroende variabeln, examensgraden, är binär och det finns ett flertal oberoende variabler tillämpas multipel logistisk regression.

Det finns ett antal grundläggande modellantaganden vid logistisk regression vilka är att rekommendera att ta hänsyn till. Dessa är följande: Endast väsentliga variabler bör tas med i modellen. Observationerna är oberoende och mäts utan fel. De oberoende variablerna ska inte ha starka samband med varandra. Sannolikheterna är en logistisk funktion av de förklarande variablerna (UCLA: Statistical Consulting Group, 2021a).

I modellen beräknas sannolikheten för att händelsen examen ska inträffa givet kännedom av värdena på de förklarande variablerna och parametrarna. Händelsen skildras av den binära variabeln y som antar värdet 1 om en individ tar ut en examen, annars 0. Den sökta sannolikheten $P(y=1|x_1,...,x_k)$ är en funktion av de förklarande variablerna och denna betecknas som $\pi(x_1,...,x_k)$. Funktionen utgör även väntevärdet av Y (Sheater, 2009):

$$\begin{aligned} E(y) &= E(y \mid x_1, \ldots, x_k) = 1 \cdot P(y = 1 \mid x_1, \ldots x_k) + 0 \cdot E(y = 0 \mid x_1, \ldots x_k) \\ &= P(y = 1 \mid x_1, \ldots, x_k) = \pi(x_1, \ldots, x_k). \end{aligned} \tag{1}$$

Hädanefter används beteckningen $\pi(\mathbf{x})$ som en kortare beteckning för $\pi(x_1,...,x_k)$, dvs $\mathbf{x}$ betecknar då ett flertal variabler. Denna sannolikhet kan endast anta värden mellan 0 och 1 (Sheater, 2009).

En multipel logistisk regression har inte ett linjärt samband mellan sannolikheten och de oberoende variablerna. Istället används logaritmen av oddset för en händelse för att beskriva en linjär funktion av de förklarande variablerna. Oddset är sannolikheten, $\pi(\mathbf{x})$, att händelsen inträffar dividerat med sannolikheten, $(1-\pi(\mathbf{x}))$, att händelsen inte inträffar,

$$\Omega = \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})}. \tag{2}$$

Ett odds på 1.0 innebär att sannolikheten för att händelsen inträffar är lika stor som att den inte gör det, det vill säga $\pi(\mathbf{x}) = 0.5$. Ett odds större än 1.0 innebär att det är mer troligt att en händelse inträffar än att den ej gör det, det vill säga $\pi(\mathbf{x}) > 0.5$. För ett odds mindre än 1.0 är det därför mindre troligt att händelsen inträffar än att den inte gör det, dvs $\pi(\mathbf{x}) < 0.5$. Oddset kan anta värden mellan 0 och oändlighet, och har därför inte samma gränsvärden som sannolikheten $\pi(\mathbf{x})$ (Allison, 2012). Den logistiska regressionsmodellen ges av uttrycket

$$\log\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k\,. \tag{3}$$

där α är interceptet, det vill säga log-oddset för att ta examen när de förklarande variablerna är noll. Vidare anger β_k hur mycket log-oddset för händelsen "y=1" förändras när x_k ökar med en enhet och de övriga variablerna hålls konstanta.

Genom att lösa ut $\pi(\mathbf{x})$ ur ekvationen räknas regressionskoefficienterna om till sannolikheter.

$$\pi(\mathbf{x}) = \frac{1}{1+\exp\left(-\alpha-\beta_1 x_k-\beta_2 x_2-\ldots-\beta_k x_k\right)} \tag{4}$$

Ekvationen har egenskapen att för alla värden på β samt *x* är sannolikheten alltid inom intervallet 0 till 1(Allison, 2012).

SAS genererar oddskvoter för samtliga kategorier inom en given variabel. Oddskvoter används för att undersöka hur en variabel förhåller sig till en annan variabel, i det här fallet, hur en kategori förhåller sig till en annan kategori. Det beräknas genom att ta kvoten av två odds,

$$Oddskvot = \frac{\Omega_1}{\Omega_2}\,. \tag{5}$$

Här anger Ω_1 respektive Ω_2 oddsen för en händelse vid två olika uppsättningar av värden, exempelvis $(x_1,...,1)$ respektive $(x_1,...,0)$. Sålunda anges hur värdet på variabeln x_k skiljer sig åt med en enhet medan de övriga variablerna antar samma värden i de två situationerna.

Notera att koefficienten β_k kan definieras som

$$\beta_k \;=\; \log \Omega_1 \;-\; \log \Omega_2 \;=\; \log\left(\frac{\Omega_1}{\Omega_2}\right). \tag{6}$$

Genom att applicera den inversa funktionen av logaritmen - exponenten - på alla led fås oddskvoten:

$$e^{\beta_k} \;=\; e^{\log\left(\frac{\Omega_1}{\Omega_2}\right)} \;=\; \frac{\Omega_1}{\Omega_2} \tag{7}$$

(UCLA: Statistical Consulting Group, 2021b).

3.2 Multikolinjäritet

Multikolinjäritet innebär att två eller fler av de förklarande variablerna har ett starkt linjärt samband. Ett problem vid multikolinjäritet är att det blir svårt att urskilja effekterna av de förklarande variablerna på den beroende variabeln. Ytterligare problem är att standardfelen för de påverkade koefficienterna blir mer benägna att vara stora; vilket leder till svårigheter vid hypotestest av koefficientens signifikans. Risken är då stor att man misslyckas med att avvisa en falsk nollhypotes, ett typ-II fel (UCLA: Statistical Consulting Group, 2021a).

3.3 Cramérs V

För att undersöka om det finns multikolinjäritet bland de förklarande variablerna i datamaterialet tillämpas Cramérs V. Metoden antar att variablerna är kategoriska. Således uppfyller majoriteten av variablerna antagandet och det är endast kategoriska variabler som kommer testas mot varandra.

Cramérs V beräknar korrelationen i tabeller som är större än 2x2 rader och kolumner. Först genomförs ett chi-två-test för att avgöra om det finns ett signifikant samband mellan variablerna. Nollhypotesen förkastas om det föreligger ett samband mellan variablerna. För att få reda på hur starkt sambandet är används följande formel (Changingminds, 2021):

$$V = \sqrt{\frac{\chi^2}{n(k-1)}}. \tag{8}$$

Här utgör V värdet för Cramérs V, χ^2 är Pearsons chi-2-test, n är stickprovsstorleken och k är minsta av antalet rader eller kolumner bland de två variablerna i en korstabell.

Cramérs V kan anta värden mellan 0 och 1. Värden nära noll tyder på ett svagt samband medan värden nära 1 indikerar ett starkt samband mellan variablerna. Bland de variabler som har en stark korrelation tas den ena variabeln bort då medtagandet av den bidrar till en sämre skattning för modellen. Dessutom tillför den då inte någon ytterligare förklaring för variationen av Y (Changingminds, 2021.)

I regel finns det tre grupper av värden man kan få fram med Cramérs V (Acastat, 2015);

- $V \in [0.0,0.3]$ - svag samvariation mellan variablerna
- $V \in [0.4,0.5]$ - mellan stark samvariation mellan variablerna
- $V > 0.5$ stark samvariation mellan variablerna

3.4 Val av modell - Stepwise selection

Vid val av vilka och antalet variabler som ska inkluderas i modellen finns det några olika riktlinjer. I regel minskar bias när antalet variabler ökar men då ökar även variansen i modellen och därmed effektiviteten av skattningen (Sheater, 2009).

För att finna den bästa modellen kommer stepwise selection att tillämpas. Tillvägagångssättet för en stepwise selection är följande: Till en början består modellen endast utav interceptet, följaktligen finns det då inga förklarande variabler inkluderade i modellen.

För att en förklarande variabel ska inkluderas i modellen får den inte överstiga det givna p-värdet som är satt till 0.05. I varje steg inkluderas den variabel som har lägst p-värde av de kvarstående variablerna. Dessutom undersöker stegvis regression i varje steg om variablerna som är inkluderade i modellen fortfarande är signifikant skilda från noll. De variabler som inte uppfyller kravet avlägsnas från modellen. Detta händelseförlopp fortsätter tills alla variabler som uppnår kriterierna har tagits med (Sheater, 2009).

3.5 Kriterier för att jämföra modeller

3.5.1 -2 log Likelihood.

Testet -2 log likelihood tar det maximala värdet av logaritmen av likelihoodfunktionen multiplicerat med -2:

$$-2 \cdot \log L \tag{9}$$

Loglikelihooden multipliceras med *-2* för att omvandla loglikelihooden till en chi-2 fördelning. Chi-2 fördelningen är viktig då den senare kan användas för att fastställa statistisk signifikans (ReStore, 2011).

Testet används dels för att undersöka hur väl anpassad modellen är till datan och dels för att jämföra modeller mot varandra. Ju högre värden desto sämre anpassning har modellen till datan. Notera dock att detta värdet påverkas signifikant av antalet observationer. Ett

problem med - 2 log L är att modeller med fler variabler har en tendens att ge lägre värden till följd av slumpen. Detta är AIC respektive BIC bättre på att åtgärda vilka redogörs för i avsnitt 3.5.2 samt 3.5.3 (Allison, 2012).

Differensen mellan två modellers - 2 log L värden, där den ena modellen är en restriktion av den andra, ger likelihoodkvot-testet. Om differensen är av avsevärd magnitud kan det indikera att de variabler som ej är inkluderade i den restriktiva modellen har en signifikant påverkan på den beroende variabeln (Pawitan, 2001).

3.5.2 Akaike Information criterion (AIC)

AIC är ett mått på hur väl anpassad modellen är till datan. Den är användbar då den tar hänsyn till goodness of fit och då den bestraffar mer komplexa modeller. Det syns i beräkningen där AIC bestraffar på k, antalet parametrar, genom:

$$AIC = -2\log L + 2k. \tag{10}$$

Om två modeller har samma förklaringsgrad, kommer modellen med färre antal parametrar att generera ett lägre AIC-värde. Den modellen är då bättre anpassad till datan eftersom modeller med lägst AIC-värde är att föredra (Allison, 2012).

3.5.3 Bayesian Information criterion - BIC

BIC ger mindre komplexa modeller än AIC då den har en hårdare bestraffning på modellens komplexitet. Även för BIC gäller det att modellen med lägst värde är att föredra. Denna beräknas genom

$$BIC = -2\log L + k\log(n) \tag{11}$$

där n är stickprovsstorlek och k är antalet parametrar (Allison, 2012).

3.6 AUC - ROC

ROC- kurvan, receiver operating characteristic curve, åskådliggör den prediktiva förmågan hos ett binärt klassificeringssystem i en plot. Kurvan skapas genom att plotta *sensitiviteten* som en funktion av *1-specificiteten* för alla möjliga brytpunkter mellan noll och ett. Sensitiviteten hänvisar till sannolikheten att händelse "ŷ=1" inträffar i modellen när händelse "y=1" är det korrekta utfallet. Specificitet är sannolikheten att händelse "ŷ=0" inträffar i modellen när händelse "y=0" är det korrekta utfallet. Därmed åskådliggör 1-specificiteten de utfall där modellen har predikterat "ŷ=1" när "y=0" är det korrekta utfallet, dvs där modellen har gjort en felaktig prediktion av det verkliga utfallet (Agresti, 2013).

Den prediktiva förmågan mäts genom att beräkna arean under kurvan, AUC. Ju större area desto bättre är modellens förmåga att förutse det korrekta utfallet. I regel kan man dela in värdena för AUC i fyra grupper:

AUC = 0.5 innebär att modellen saknar prediktiv förmåga.

$0.7 \leq \text{AUC} \leq 0.8$ innebär att modellens prediktiva förmåga är godtagbar.

$0.8 \leq \text{AUC} \leq 0.9$ innebär att modellens prediktiva förmåga är god.

$0.9 \leq \text{AUC}$ innebär att modellens prediktiva förmåga är utmärkt.

(Hosmer&Lemeshow, 2000)

4. Resultat

4.1 Kontroll av multikolinjäritet

I det här avsnittet bedöms hur starkt sambandet mellan de förklarande variablerna är med hjälp av måttet Cramérs V. Det är endast de förklarande variabler som författaren har spekulerat om att det kan föreligga ett samband mellan som testas.

Tabell 3: Kontroll av multikolinjäritet mellan ett urval av de oberoende variablerna

Variabler som testas	**Cramérs V**
utlsSvBakg_YH * Inrikes_utrikes_yh	1
Utbildningsnivå * föräldrarnas_Utb_nivå	0.1638
Utbildningsnivå * jmftal_grupperad (betyg)	0.2373
Kön * Utbildningsområde	0.5407
Sysselsättning_före * Sysseslsättning_under	0.5517
jmftal_grupperad * Behor_sök_plats_grupperad	0.0549
studieform * sysselsättning_före	0.1177
Kön * Jmftal_grupperad (betyg)	0.1947
Utbildningsnivå* Inrikes_utrikes_yh	0.2207
Behor_sök_plats_grupperad * Utbildningsområde	0.1884
Utbildningsnivå * Utbildningsområde	0.1084
studieform * Utbildningsområde	0.3342

UtlSvBakg_YH och *Inrikes_utrikes_yh* kontrollerades då den förstnämnda variabeln är en mer detaljerad klassificering av Inrikes_utrikes_yh. Värdet för Cramérs V blev *1* vilket innebär att de har ett perfekt samband. Vid skapandet av modellen testades båda variabler var för sig. Den variabel som bedömts ha det starkaste sambandet med examensgraden var *Inrikes_utrikes_yh* och därför används endast den framöver.

Cramérs V värdet för *Kön* och *Utbildningsområde* blev 0.5407. Sålunda fanns det en stark samvariation mellan de två förklarande variablerna. Det har från den deskriptiva statistiken observerats att en större andel kvinnor tar examen än män, därför är båda variabler av intresse. För den skull skapades senare två modeller, en där *Kön* är inkluderad och en där *Kön* är exkluderad.

Slutligen testades *Sysselsättning_före* mot *Sysselsättning_under* eftersom det verkade troligt att en individ som haft en sysselsättning före utbildningen även skulle ha det under utbildningen. Cramérs V värdet blev *0.5517* vilket innebar att variablerna hade en hög samvariation. Den variabel som hade det starkaste sambandet med examensgraden var *Sysselsättning_före* och därför används endast den framöver.

4.2 Sökandet efter en lämplig modell

Vid sökandet efter de variabler som bäst förklarade variationen i examensgraden användes främst stepwise regression. Men innan den implementerades sorterades variabler som exempelvis radnummer, utbildningsnummer och omgångsnummer bort från datamaterialet. Detta då de ej bedöms ha någon verklig effekt på examensgraden. De variabler som hade ett starkt samband har även de tagits bort vid sökandet efter en passande modell. Det fanns även tre olika uppsättningar av variabler fördelat på år som relaterade till individens sysselsättning före och under utbildningen. Dessa sorterades och togs bort. Istället kodades två nya variabler *Sysselsättning_före* och *Sysselsättning_under*. I dessa variabler har sysselsättning betingats på individen och således är variablerna inte längre beroende av året för sysselsättning.

Därefter genomfördes stepwise regression ett antal gånger, dels med alla variabler inkluderade och med kombinationer av olika variabler. De variabler som ständigt återkom i modellerna och som hade högst signifikans var följande: *Utbildningsområde*, *jmftal_grupperad(betyg)*, *studieform*, *utbildningsnivå*, *Sysselsättning_före*, *Inrikes_utrikes_yh*, *Behor_sök_plats_grupperad (behöriga sökande per plats indelat i 8 grupper*) samt *Kön*.

Trots att *Kön* och *Utbildningsområde* hade en relativt hög samvariation togs båda med i modell 1. Detta eftersom den deskriptiva statistiken visat att andelen examinerade var lägre för män än för kvinnor inom de de flesta utbildningsområden. *Kön* bedömdes därför som en viktig förklaringsfaktor till examensgraden. Men då de samvarierar kan det vara svårt att urskilja variablernas verkliga effekt på den beroende variabeln. I denna tillämpning är dock datamaterialet stort, vilket innebär en relativt liten varians för samtliga punktskattningar i de valda modellerna. Därför bedöms inte multikolinjäriteten vara ett utmärkande problem i modellen.

Modell 1:

Inkluderar variablerna *Kön*, *Utbildningsområde*, *studieform*, *sysselsättning_före*, *Inrikes_utrikes_yh*, *jmftal_grupperad(betyg)*, *Behor_sök_plats_grupperad* och *Utbildningsnivå*.

Tabell 4: Oddskvoter tillhörande modell 1

Effekt	Punkt-skattning	95 % Wald	Konfidens-intervall
Jmftal_grupperad 00,0-09,99 vs 17,5-20,00	0.430	0.389	0.475
Jmftal_grupperad 10,0-12,49 vs 17,5-20,00	0.591	0.536	0.651
Jmftal_grupperad 12,5-14,99 vs 17,5-20,00	0.833	0.756	0.917
Jmftal_grupperad 15,0-17,49 vs 17,5-20,00	1.031	0.931	1.141
Jmftal_grupperad uppgift saknas vs 17,5-20,00	0.576	0.524	0.633
Behor_sök_plats_grupperad 02,0-02,99 vs 00,0-01,99	1.173	1.127	1.221
Behor_sök_plats_grupperad 03,0-03,99 vs 00,0-01,99	1.324	1.256	1.395
Behor_sök_plats_grupperad 04,0-04,99 vs 00,0-01,99	1.459	1.362	1.563
Behor_sök_plats_grupperad 05,0-05,99 vs 00,0-01,99	1.667	1.543	1.800
Behor_sök_plats_grupperad 06,0-06,99 vs 00,0-01,99	1.493	1.354	1.646
Behor_sök_plats_grupperad 07,0-07,99 vs 00,0-01,99	1.419	1.258	1.601
Behor_sök_plats_grupperad 8 + vs 00,0-01,99	1.723	1.589	1.868
Behor_sök_plats_grupperad uppgift saknas vs 00,0-01,99	1.206	1.148	1.266
Utbildningsområde Ekonomi, administration och försäljning vs Data/IT	1.884	1.783	1.990
Utbildningsområde Friskvård och kroppsvård vs Data/IT	1.471	1.196	1.808
Utbildningsområde Hotell, resturang och turism vs Data/IT	1.402	1.293	1.520
Utbildningsområde Hälso- och sjukvård samt socialt arbete vs Data/IT	2.363	2.214	2.521
Utbildningsområde Journalistik och Information vs Data/IT	1.450	1.280	1.642
Utbildningsområde Juridik vs Data/IT	1.084	0.913	1.287
Utbildningsområde Kultur, media och design vs Data/IT	2.346	2.159	2.548
Utbildningsområde Lantbruk, djurvård, trädgård, skog och fiske vs Data/IT	1.799	1.630	1.986
Utbildningsområde Pedagogik och undervisning vs Data/IT	3.721	3.286	4.214
Utbildningsområde Samhällsbyggnad och byggteknik vs Data/IT	1.446	1.367	1.528
Utbildningsområde Säkerhetstjänster vs Data/IT	2.652	2.225	3.160
Utbildningsområde Teknik och tillverkning vs Data/IT	1.310	1.241	1.382
Utbildningsområde Transporttjänster vs Data/IT	2.103	1.916	2.307
Utbildningsområde Övrigt vs Data/IT	1.859	1.516	2.279
Sysselsättning_före Sysselsatt vs Ej sysselsatt	1.646	1.591	1.702
Sysselsättning_före Uppgift saknas vs Ej sysselsatt	1.588	1.343	1.877
Kön 1 vs 2	0.752	0.726	0.779
Utbildningsnivå Förgymnasial vs Eftergymnasial	0.760	0.689	0.837
Utbildningsnivå Gymnasial vs Eftergymnasial	1.107	1.071	1.146
Utbildningsnivå Uppgift saknas vs Eftergymnasial	1.161	0.953	1.415
Inrikes_Utrikes_yh Inrikes född vs Utrikes född	1.393	1.340	1.447
Inrikes_Utrikes_yh Uppgift saknas vs Utrikes född	0.240	0.121	0.478
Studieform Bunden vs Distans	2.252	2.170	2.337

I tabell 4 presenteras oddskvoterna samt konfidensintervallen mellan de förklarande variablernas kategorier. Oddskvoterna kan tolkas på följande sätt: Effektkolumnen återger de kategorier inom variabeln som har testats mot varandra. Kategorin till vänster i effektkolumnen är kategorin som testas mot referensgruppen som finns till höger. Till exempel visas på första raden oddskvoten för variabeln *jmftal_grupperad(betyg),* till vänster finns betygsgruppen *00,0-09,99* och till höger referensgruppen *17,5-20,00.* I kolumnen för punktskattningar ser man att värdet på oddskvoten är 0.430. Värdet som genereras visar oddset för gruppen *00,0-09,99* i relation till referensgruppen *17,5-20,00.* Vidare innebär det att gruppen *00,0-09,99* har lägre sannolikhet att ta examen än referensgruppen *17,5-20,00.* Det kan även konstateras att samtliga betygsgrupper förutom betygsgrupp *15,0-17,49* har lägre sannolikhet att ta examen än referensgruppen *17,5-20,00.* Det innebär att individer med högre gymnasiebetyg har större sannolikhet att ta examen än individer med lägre gymnasiebetyg.

Konfidensintervallet för oddskvoten mellan betygsgruppen *00,0-09,99* och referensgruppen *17,5-20,00* visar att värdet befinner sig mellan 0.389 och 0.475. Den exakta innebörden av konfidensintervallet kan vara svårt att tolka. Dock kan det konstateras att det finns en signifikant skillnad mellan grupperna då intervallet inte täcker 1.0. Konfidensintervall som täcker värdet 1.0 innebär att det inte finns en signifikant skillnad mellan grupperna. Detta eftersom att en oddskvot på 1.0 innebär att sannolikheten för att händelsen inträffar är lika stor som att den inte gör det, det vill säga sannolikheten är lika med 0.50. Vidare innebär det att det inte kan påvisas någon skillnad mellan sannolikheten för kategorin att ta examen jämfört med referensgruppen. Konfidensintervallet mellan betygsgrupp *15,0-17,49* och referensgrupp *17,5-20,00* är mellan 0.931 och 1.141. Det innebär att det inte kan påvisas någon skillnad i sannolikheten att ta examen mellan grupperna.

Oddskvoterna för variabeln *Behor_sök_plats_grupperad* visar att individer som påbörjat utbildningar med fler behöriga sökande per plats, är mer troliga att ta examen än de individer som påbörjat en utbildning med färre behöriga sökande per plats. Till exempel är oddskvoten mellan den lägsta Behor_sök_plats_grupperad - gruppen (*00,0-01,99*) och den högsta Behor_sök_plats_grupperad - gruppen (*8+*) 1.723.

Oddskvoten för variabeln *Utbildningsområde* jämför data/IT med övriga utbildningsområden. Resultatet visar att en individ som studerar data/IT är mindre trolig att ta examen i jämförelse med

de flesta andra utbildningsområden. Oddskvoten mellan data/IT och juridik har ett värde på 1.084. Konfidensintervallet befinner sig mellan 0.913 och 1.287. Intervallet täcker sålunda 1.0 vilket innebär att det inte kan påvisas någon skillnad i sannolikheten att ta examen mellan grupperna.

Oddskvoten för variabeln *Sysselsättning_före* jämför sysselsatt med ej sysselsatt och har ett värde på 1.646. Det innebär att en individ som varit sysselsatt före påbörjad utbildningsomgång har högre sannolikhet att ta examen än en individ som inte varit det.

Oddskvoten för *Kön* jämför män med kvinnor, värdet blev 0.752 vilket innebär att en kvinna har större sannolikhet att ta examen än en man.

För variabeln *Utbildningsnivå* är det främst oddskvoten mellan förgymnasial- och eftergymnasial utbildning som är utmärkande. Oddskvoten är 0.76 vilket innebär att en individ med en eftergymnasial utbildning är mer trolig att ta examen än en individ med enbart förgymnasial utbildning. Notera också att oddskvoten mellan gymnasial och eftergymnasial har ett värde på 1.107. Det innebär att en individ som endast har en gymnasial utbildning har något högre sannolikhet att ta examen än en individ som har en eftergymnasial utbildning.

Oddskvoten för variabeln *Inrikes_utrikes_yh* har ett värde på 1.393. Det innebär att det är mer troligt att en inrikes född individ tar examen än att en utrikes född individ gör det. Den sista raden i tabell 4 presenterar oddskvoten för variabeln *studieform* som har värdet 2.252. Det indikerar att det är mer troligt att ta examen från en bunden utbildning jämfört med en utbildning som bedrivs på distans.

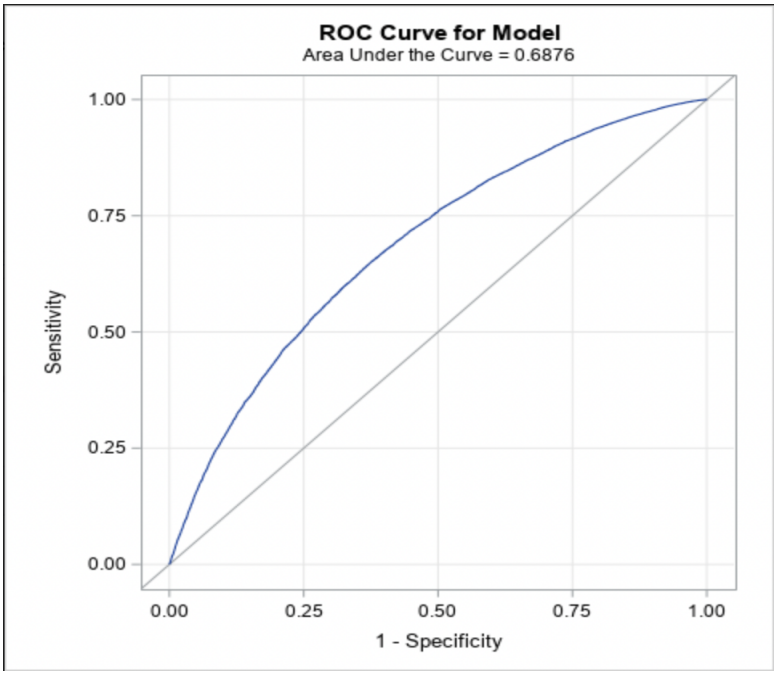

Diagram 1: ROC-kurva för modell 1

Diagram 1 illustrerar med hjälp utav ROC-kurvan modell 1:s prediktiva förmåga. Värdet för arean under kurvan är 0.6876. Vidare är det ett värde som är mycket nära 0.7 och modellens prediktionsförmåga anses därför vara godtagbar (se avsnitt 3.6).

Modell 2:

I modell 2 exkluderades variabeln *Kön* eftersom den samvarierar med variabeln *Utbildningsområde*. Modellen inkluderade därefter variablerna *Utbildningsområde*, *studieform*, *Sysselsättning_före*, *Inrikes_utrikes_yh*, *jmftal_grupperad(betyg)*, *Behor_sök_plats_grupperad* och *Utbildningsnivå*. Följande resultat genererades:

Tabell 5: Oddskvoter tillhörande modell 2

Effekt	Punkt-skattning	95 % Wald	Konfidens-intervall
Jmftal_grupperad 00,0-09,99 vs 17,5-20,00	0.402	0.363	0.444
Jmftal_grupperad 10,0-12,49 vs 17,5-20,00	0.548	0.498	0.604
Jmftal_grupperad 12,5-14,99 vs 17,5-20,00	0.790	0.717	0.870
Jmftal_grupperad 15,0-17,49 vs 17,5-20,00	1.004	0.907	1.111
Jmftal_grupperad uppgift saknas vs 17,5-20,00	0.555	0.504	0.610
Behor_sök_plats_grupperad 02,0-02,99 vs 00,0-01,99	1.181	1.135	1.229
Behor_sök_plats_grupperad 03,0-03,99 vs 00,0-01,99	1.341	1.272	1.413
Behor_sök_plats_grupperad 04,0-04,99 vs 00,0-01,99	1.490	1.391	1.597
Behor_sök_plats_grupperad 05,0-05,99 vs 00,0-01,99	1.697	1.571	1.832
Behor_sök_plats_grupperad 06,0-06,99 vs 00,0-01,99	1.522	1.381	1.678
Behor_sök_plats_grupperad 07,0-07,99 vs 00,0-01,99	1.433	1.270	1.617
Behor_sök_plats_grupperad 8 + vs 00,0-01,99	1.756	1.619	1.904
Behor_sök_plats_grupperad uppgift saknas vs 00,0-01,99	1.208	1.150	1.268
Utbildningsområde Ekonomi, administration och försäljning vs Data/IT	2.156	2.046	2.272
Utbildningsområde Friskvård och kroppsvård vs Data/IT	1.785	1.454	2.192
Utbildningsområde Hotell, resturang och turism vs Data/IT	1.626	1.503	1.759
Utbildningsområde Hälso- och sjukvård samt socialt arbete vs Data/IT	2.835	2.668	3.013
Utbildningsområde Journalistik och Information vs Data/IT	1.633	1.444	1.847
Utbildningsområde Juridik vs Data/IT	1.279	1.079	1.516
Utbildningsområde Kultur, media och design vs Data/IT	2.501	2.304	2.716
Utbildningsområde Lantbruk, djurvård, trädgård, skog och fiske vs Data/IT	2.014	1.827	2.221
Utbildningsområde Pedagogik och undervisning vs Data/IT	4.095	3.617	4.636
Utbildningsområde Samhällsbyggnad och byggteknik vs Data/IT	1.462	1.383	1.545
Utbildningsområde Säkerhetstjänster vs Data/IT	2.773	2.328	3.304
Utbildningsområde Teknik och tillverkning vs Data/IT	1.306	1.238	1.379
Utbildningsområde Transporttjänster vs Data/IT	2.160	1.969	2.369
Utbildningsområde Övrigt vs Data/IT	2.053	1.676	2.515
Sysselsättning_före Sysselsatt vs Ej sysselsatt	1.651	1.597	1.708
Sysselsättning_före Uppgift saknas vs Ej sysselsatt	1.602	1.355	1.893
Utbildningsnivå Förgymnasial vs Eftergymnasial	0.725	0.658	0.799
Utbildningsnivå Gymnasial vs Eftergymnasial	1.090	1.054	1.128
Utbildningsnivå Uppgift saknas vs Eftergymnasial	1.136	0.932	1.384
Inrikes_utrikes_yh Inrikes född vs Utrikes född	1.401	1.348	1.456
Inrikes_utrikes_yh Uppgift saknas vs Utrikes född	0.246	0.124	0.489
Studieform_utb Bunden vs Distans	2.248	2.166	2.332

Resultatet i tabell 5 är snarlikt resultatet i tabell 4. Till exempel visas på första raden oddskvoten för variabeln *jmftal_grupperad(betyg),* till vänster finns betygsgruppen *00,0-09,99* och till höger referensgruppen *17,5-20,00.* I kolumnen för punktskattningar ser man att värdet på oddskvoten är 0.402. Motsvarande värde för modell 1 i tabell 4 är 0.430. Skillnaden mellan oddksvoterna från tabell 5 och tabell 4 syns främst när man studerar variabeln *Utbildningsområde*. När *Kön* exkluderas från modellen ökar oddskvoterna. I tabell 4 har utbildningsområdet ekonomi, administration och försäljning ett värde på 1.884 när det jämförs med data/IT. I tabell 5 är motsvarande värde 2.156. En viss ökning kan alltså påvisas vilket beror på att variabeln *Utbildningsområde* får en större påverkan i modellen när *Kön* exkluderas.

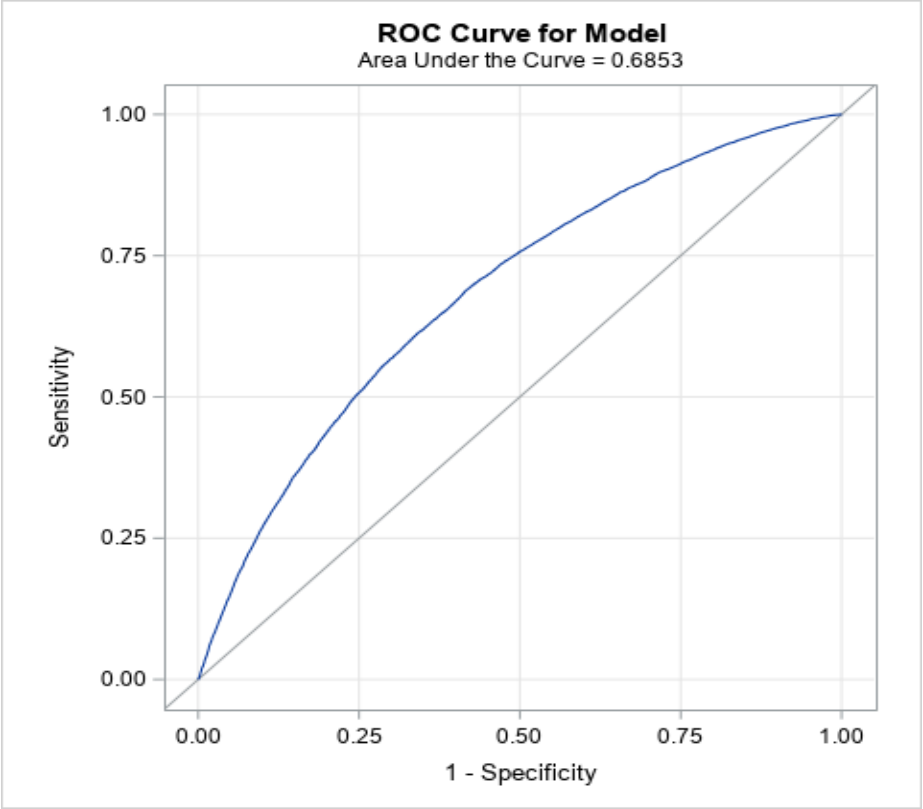

Diagram 2: ROC-kurva för modell 2

Diagram 2 illustrerar den prediktiva förmågan hos modell 2, nämligen hur bra modellen är på att förutspå utfallet examen när examen är det korrekta utfallet. Värdet på arean under kurvan är 0.6853, det vill säga något lägre än modell 1 men den anses fortfarande vara godtagbar (se avsnitt 3.6).

Tabell 6: Kriterier för att avgöra vilken modell som är mest lämplig

Modell	**1**	**2**
AIC	107 271.43	107 518.68
BIC	107 622.88	107 860.63
-2 log L	107 197.43	107 446.68

I tabell 6 visas värdena för de tre testerna, AIC, BIC samt -2 log L, som används för att avgöra vilken modell som är mest lämplig, se avsnitt 3.5. Resultaten visar att modell 1 har genererat de lägsta värdena inom samtliga kriterier, och anses därför vara den bästa modellen av de två. Differensen mellan - 2 log L värdet för modell 2 och modell 1 är 249.25. Differensen är approximativt chi-2(1)-fördelad med en avsevärd magnitud och indikerar att variabeln *Kön* har en signifikant påverkan på utfallet examen, (p-värdet = 0.0000).

5 Diskussion

I resultatdelen presenterades två modeller, en där variabeln *Kön* var inkluderad och en där den var exkluderad. I båda modellerna förekom variablerna *Utbildningsområde*, *studieform*, *jmftal_grupperad(betyg)*, *Inrikes_utrikes_yh*, *Behor_sök_plats_grupperad*, *Utbildningsnivå* samt *Sysselsättning_före*. Två modeller skapades eftersom *Utbildningsområde* och *Kön* samvarierar men anses båda bidra till förklaringen av variationen i examensgraden. När man jämför resultatet för de två modellerna visade det sig att *Utbildningsområdets* oddskvoter ökade när *Kön* exkluderades. Det beror på att *Utbildningsområde* då får en större påverkan på examensgraden eftersom effekten av *Kön* tas bort från modellen. Det är samspelseffekten mellan *Kön* och *Utbildningsområde* som ger uttryck, vilket synliggörs i skillnaden mellan parameter skattningarna för de olika modellerna. Sålunda bidrar *Kön* med en del information, men eftersom det finns multikolinjäritet är det svårt att avgöra hur mycket respektive variabel påverkar examensgraden. Däremot är det ett stort datamaterial vilket medför att skattningarna, trots förekomst av multikolinjäritet, har relativt liten varians. Därmed bedöms det motiverat att ha med båda variablerna som bidragande faktorer till variationen i examensgraden.

Tabell 6 visar att värdena för AIC, BIC och - 2 log L skiljer sig åt något mellan modellerna. Den modell som genererade de lägsta värdet var modell 1. Således visar informationen att modell 1 är mest väl anpassad till datan. Testerna visar därmed att *Kön* har en signifikant effekt på examensgraden. Likelihoodkvot-testet indikerar också att *Kön* har en signifikant påverkan på examensgraden eftersom differensen mellan modellernas - 2 log likelihood värden är approximativt chi-2(1)-fördelad med en betydande magnitud. ROC-kurvan för modell 1 ger ett tämligen högre värde än för modell 2. Det innebär att prediktionsförmågan i modell 1 är bättre än i modell 2, precis som testerna visade. Inkluderandet av variabeln *Kön* bidrar därför till viss del med att predicera det korrekta utfallet. Därmed kan det konstateras att både *Utbildningsområde* och *Kön* har en viss påverkan på examensgraden, hur mycket och vilken som är mest betydande hade kunnat undersökas vidare genom att studera samma mönster vid andra eftergymnasiala utbildningar.

I den här delen spekuleras det kring varför variablerna i modellerna kan tänkas samvariera med examensgraden. I tabell 4 och 5 presenterades oddskvoterna för modell 1 respektive 2. Från dem kunde man bland annat avläsa att individer med högre betyg har större chans att ta examen än individer med lägre betyg. Högre betyg kan indikera att det är en person som lägger ner mycket tid på sina studier. Det verkar därför troligt att individen är mer motiverad att fullfölja sin utbildning. En individ med en gymnasial eller eftergymnasial utbildning är också mer trolig att ta examen. Det kan bero på att individen har mer studieerfarenhet och därför har bättre förutsättningar att fullfölja sina studier. Ytterligare en tänkbar möjlighet är att en individ som studerar på yrkeshögskola efter att ha fullföljt en gymnasial eller eftergymnasial utbildning gör det för att utvecklas inom sitt yrke, exempelvis för att få nya arbetsuppgifter och bättre lönevillkor. Vilket kan fungera som ett incitament till att fullfölja utbildningen.

Individer som har varit sysselsatta före påbörjad utbildning har större sannolikhet att ta examen än individer som inte har haft en sysselsättning innan. En möjlig förklaring är att individer som har haft en sysselsättning inom samma eller relaterade branscher både besitter användbara förkunskaper samt större intresse för sitt utbildningsval. Det kan fungera som ett

incitament att fullfölja utbildningen. Det kan även vara motsatsen, nämligen att man har en sysselsättning man inte trivs med och därför vill studera vidare för att byta inriktning.

De utbildningar med fler behöriga sökande per plats har högre examensgrad än de utbildningar med färre behöriga sökande per plats. Det kan bero på att de förstnämnda utbildningarna är kända för att ge goda möjligheter efter avslutad utbildning. Ytterligare en anledning kan vara att ett högre söktryck gör att de individer som blir antagna blir mer motiverade till studier då det är en eftertraktad utbildning.

En individ som påbörjat en utbildning på plats är mer trolig att ta examen än en individ som påbörjat en utbildning på distans. När utbildningen bedrivs på plats skapas en bättre gemenskap mellan andra studenter, studenten har en plats för studier och kommunikationen med lärare i och utanför klassrummet blir enklare. En individ som är inrikes född är mer trolig att ta examen än en individ som är utrikes född. Det kan exempelvis bero på att en inrikes född individ besitter mer kunskaper i det svenska språket eller har en mer etablerad tillvaro vilket underlättar fullföljandet av studierna.

Det är dock viktigt att notera att detta endast är spekulationer kring orsakssambanden mellan examensgraden och de förklarande variablerna. Anledningarna till studieavhopp är många och påverkas av flera olika faktorer som interagerar med varandra. Att fastställa exakta orsakssamband kräver därför en mer komplex och avancerad analys. Framförallt eftersom det i rapporten endast kan konstateras att det föreligger ett statistiskt samband mellan examensgraden och variablerna i modellerna, men inte konstateras exakt varför.

6 Slutsats

Syftet med rapporten var att fastställa vilka faktorer som samvarierar med examensgraden vid yrkeshögskolan. Utifrån analys kom författaren fram till att de viktigaste förklarande variablerna är *Kön, studieform*, *jmftal_grupperad (betyg)*, *Utbildningsområde*, *Sysselsättning_före*, *Inrikes_utrikes_yh*, *Behor_sök_plats_grupperad* samt *Utbildningsnivå*. Studien visar att individer som besitter följande egenskaper har en ökad sannolikhet att ta examen: Studerade med en bunden utbildning, inte ett tekniskt utbildningsområde, inrikes

född, kvinna, hade en sysselsättning före påbörjad utbildning, betyg i någon utav de högre betygs grupperna, påbörjade en utbildning med många behöriga sökande per plats samt hade en gymnasial eller eftergymnasial utbildning.

Det finns en del variation i den beroende variabeln som inte går att predicera med hjälp utav de variabler som finns i det givna datamaterialet. I framtiden hade det exempelvis varit intressant att undersöka hur yrkeshögskolans geografiska plats påverkar examensgraden. Staden studierna bedrivs i påverkar individens studentliv och kan därmed påverka examensgraden. Andra faktorer som hade varit intressanta att studera är skolans utformning, pedagogik och lärare. Ytterligare faktorer som kan tänkas ha en påverkan på examensgraden är individens psykiska välmående vid studierna samt externa faktorer som exempelvis corona pandemin.

Referenser

Acastat (2015). Available online: http://www.acastat.com/statbook/chisqassoc.htm [Accessed 18 November 2021]

Agresti, A. (2013). Categorical data analysis. 3rd ed. Hoboken: Wiley

Allison, P.D. (2012). Logistic regressions using SAS: Theory and application. 2. ed. SAS Publishing

Changingminds (2021). Available online: http://changingminds.org/explanations/research/analysis/cramers_v.htm [Accessed 15 November 2021]

Hosmer, David W. & Lemeshow, Stanley (2000). Applied logistic regression. 2. ed. New York: Wiley

Introduction to SAS. UCLA: Statistical Consulting Group (2021a). Available online: https://stats.oarc.ucla.edu/stata/webbooks/logistic/chapter3/lesson-3-logistic-regression-diagnostics/ [Accessed 5 Oktober 2021]

Introduction to SAS. UCLA: Statistical Consulting Group (2021b). Available online: https://stats.oarc.ucla.edu/stata/faq/how-do-i-interpret-odds-ratios-in-logistic-regression/ [Accessed 14 Januari 2021]

Kang, H. (2013). The prevention and handling of the missing data, *Korean J Anesthesiol*. Vol. 64 , no. 5, pp. 402-406, Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/ [Accessed 20 November 2021]

Myndigheten för yrkeshögskolan (2021a). Statistisk årsrapport, Available online: https://assets.myh.se/docs/publikationer/statistiska-arsrapporter/statistisk-arsrapport-2021.pdf [15 Oktober 2021]

Myndigheten för yrkeshögskolan (2021b). Available online: https://www.myh.se/Om-oss/Organisation/Generaldirektoren-GD/GD-har-ordet/Examensgraden-inom-YH--atgarder-kravs/) [Accessed 3 Oktober 2021]

Pawitan, Y. (2001). In All Likelihood: Statistical Modelling and Inference Using Likelihood. Oxford: Clarendon Press

ReStore - National center for research methods (2011). Available online: https://www.restore.ac.uk/srme/www/fac/soc/wie/research-new/srme/glossary/index1695.html?selectedLetter=D#deviance-2ll [Accessed 14 Januari 2021]

Science (2021). Available online: Logistisk regression: genomförande, tolkning, odds ratio, multipel regression - Science [Accessed 5 Oktober 2021]

Sheater, S. (2009). A modern approach to Regression with R [elektronisk resurs]. New York: Springer

Statistiska centralbyrån (2021). Available online: https://www.scb.se/hitta-statistik/statistik-efter-amne/utbildning-och-forskning/eftergymnasial-yrkesutbildning/yrkeshogskolan/pong/statistiknyhet/examinerade-fran-yrkeshogskolan-2020/ [Accessed 19 January 2022]