Stockholms universitet

# En logistisk regressionsanalys av faktorer som påverkar gymnasieelevers framtida utbildningsval

Esma Özdolap

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2023:6
Matematisk statistik
Juni 2023

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2023:6**
http://www.math.su.se/matstat

# En logistisk regressionsanalys av de faktorer som påverkar gymnasieelevers framtida utbildningsval

Esma Özdolap*

Juni 2023

**Sammanfattning**

I detta arbete analyseras, via statistik från Skolverket, hur andelen som läser vidare på högskolenivå inom ett år efter avslutad gymnasieutbildning beror på olika faktorer som kön, föräldrarnas utbildningsnivå, nationell bakgrund, samt vilket år under perioden 2014–2019 de påbörjade högskolestudier. Endast elever som fått behörighet att söka vidare samt studerat ett högskoleförberedande program kommer att studeras. Datamaterialet kommer att delas upp i tre grupper för att undersöka varje grupp individuellt, detta eftersom datamaterialet inte är individbaserat samt uppdelat på alla kategorier simultant.

Syftet med arbetet är att studera en trend samt undersöka hur viljan att studera vidare efter avslutad gymnasieutbildning beror på olika bakomliggande faktorer. Målet är att hitta den modell som bäst förklarar vilka faktorer som påverkar viljan att studera vidare efter avslutad gymnasieutbildning, och för detta ändamål använder vi oss av en logistisk regressionsmodell.

I denna analys kommer vi fram till att en tillämpad logistisk regressionsmodell för åren 2014–2020 ger en positiv trend, det vill säga att den predikterade sannolikheten att studera vidare efterföljande år ökar under perioden, medan det för samma modell mellan åren 2014–2019 minskar.

Efter att ha inspekterat datamaterialet har vi identifierat en tydlig pandemieffekt. För att ta hänsyn till denna effekt har vi utvecklat en modell som inkluderar 2020 som en förklarande variabel. Denna modell visar oss att det finns en tydlig pandemieefekt, eftersom den resulterar i bättre AIC- och BIC-värden.

Analysen visar även att pandemieffekten haft en positiv påverkan på studenternas beslut att studera vidare, för både män och kvinnor inom de olika kategorierna. En möjlig förklaring till detta är eftersom effekten av pandemin också var att allt fler blev arbetslösa.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: esma2001@hotmail.se. Handledare: Daniel Ahlberg och Johannes Heiny.

## Abstract

In this paper, we analyse the proportion of students who decide to continue their studies within one year after completing high school, using data from Skolverket between the years 2014 and 2020. We further filter the data by only looking at students who are eligible to continue their studies and who studied a college preparatory program in high school. We will look at the following factors to see what impacts the decision: the student's high school educations, their genders, their parents' educations, and the students' national backgrounds.

The data will be split into three groups and each group will be studied separately. This is because our data do not contain interactions between the factors of the high school program, ethnic background and parents' education. The purpose of this paper is to study a trend as well as finding which factors influence the student's decision to continue their study.

The goal is to find a model that best describes which factors influence the student's decision to continue their study and we use logistic regression to accomplish this goal. We found a logistic regression model for the years 2014–2020 that showed a positive trend. This model also predicts the probability of students who continue to study grows in the coming years. We suspect that the pandemic influenced the students' decisions because the unemployment rate went up during this period. Therefore, we chose to do the logistic regression on the years 2014–2019 instead. This new model shows us a negative trend.

## Förord

Denna rapport utgör ett examensarbete i matematisk statistik och motsvarar 15 högskolepoäng. Arbetet är skrivet vid Matematiska institutionen på Stockholms universitet.

Jag vill framföra ett stort tack till mina handledare Daniel Ahlberg och Johannes Heiny för all deras hjälp, synpunkter och återkoppling under arbetets gång. Vidare vill jag säga ett stort tack till Ola Hössjer och Mohamed El Khalifi för deras synpunkter och hjälp.

Jag vill även tacka mina kära föräldrar för deras stöd och synpunkter under hela arbetets gång. Vidare vill jag framföra ett stort tack till alla mina klasskamrater som under detta arbete stöttat mig med sin hjälp, återkoppling och synpunkter.

Dessutom måste jag erkänna att jag använt mig av chattroboten ChatGPT för att korrigera samt konstruera R-kod i mitt arbete. Bortsett från det har jag inte använd ChatGPT för att skriva teori eller andra delar av rapporten.

# Innehåll

**1 Inledning** **1**
1.1 Bakgrund . . . 1
1.2 Mål och syfte . . . 2

**2 Teori** **3**
2.1 Generaliserad linjär modell . . . 3
2.1.1 Slumpmässig komponent . . . 3
2.1.2 Systematisk komponent . . . 3
2.1.3 Länkfunktion . . . 4
2.2 Odds samt oddskvot . . . 4
2.3 Logistisk regression . . . 5
2.3.1 Multipel logistisk regression . . . 5
2.3.2 Oddskvot i logistisk regression . . . 5
2.4 Likelihood . . . 6
2.5 Akaikes informationskriterium, AIC . . . 6
2.6 Förklaringsgrad, $R^2$ . . . 7
2.6.1 McFaddens $R^2$ . . . 7
2.7 Cohens kappa . . . 7
2.8 Bayesianska informationskriteriet, BIC . . . 8

**3 Databeskrivning** **9**
3.1 Datavisualisering . . . 10

**4 Modellering och analys** **15**
4.1 Logistisk regression . . . 16
4.2 Programvara . . . 16
4.3 Modellframtagning . . . 17
4.3.1 Modell för åren 2014–2019 . . . 18
4.3.2 Pandemiår . . . 21
4.4 Mcfaddens $R^2$-värde . . . 25
4.5 Cohens kappa . . . 25
4.6 AIC . . . 26
4.7 BIC . . . 27

**5 Diskussion och slutsats** **28**

**Referenser** **32**

**Appendix** **34**

# 1 Inledning

Skolverket har för varje år samlat in data för elever som examinerats från gymnasieskolan beroende på kön, nationell bakgrund samt föräldrarnas högsta utbildningsnivå. Med detta kan vi analysera vilka faktorer som påverkar elevers beslut att studera vidare efter avslutad gymnasieutbildning. Från analysen av datamaterialet kan vi se att andelen som studerar vidare skiljer sig signifikant beroende på program, bakgrund och föräldrars utbildning. Vidare har vi kommit fram till att den negativa trenden, innan pandemin, varit ganska svag medan det år 2020 lutade mot en positivt lutande trend, detta på grund av effekterna som ett pandemiår medfört.

## 1.1 Bakgrund

Universiteten samt högskolorna har under de senaste åren utvecklats och expanderats. Allt fler universitet och högskolor ger i dagsläget flera olika program som man kan studera på flera olika sätt. Exempelvis kan man börja studera ett kandidatprogram efter gymnasiet eller välja att studera enstaka kurser. Allt fler möjligheter har skapats för att elever ska ha en chans att studera vidare efter gymnasiet. Hur många av eleverna är det som tar till vara på denna chans? Och varför väljer nya gymnasieexaminerade elever att studera vidare?

Andelen kvinnor som väljer att studera vidare efter avslutad gymnasieutbildning dominerar jämfört med män. Amnéus och Inkine (2016, s. 42) skriver:

> Kvinnorna är i majoritet (med mer än 60 procent av högskolenybörjarna) inom juridik och samhällsvetenskap, humaniora och teologi, medicin och odontologi samt vård och omsorg medan män är i majoritet inom teknik respektive naturvetenskap. [...] Kvinnor har varit i majoritet bland studenterna i högskolan ända sedan 1977 års högskolereform, då flera kvinnodominerade utbildningar införlivades i högskolan. Innan dess var männen fler än kvinnorna.

Andelen elever som har en svensk bakgrund som studerar vidare har minskat medan de elever med en utländsk bakgrund har ökat. Söderman (2020, s. 2) skriver:

> Under sex läsår i rad har antalet högskolenybörjare ökat i den svenska högskolan. Det senaste året ökade de till totalt 93 700 högskolenybörjare eller 1 600 studenter fler än läsåret innan. [...] Antalet inresande studenter ökade med 7 000 studenter till 23 900, samtidigt minskade antalet svenska nybörjare med 5 400 studenter jämfört med läsåret innan.

Bland de elever som får en grundläggande behörighet till universitet- och högskolestudier efter avslutad gymnasieutbildning är majoriteten kvinnor. Jacobsson, et al. (2008, s. 16) skriver:

> Kvinnor har i högre utsträckning än män grundläggande behörighet

till universitets- och högskolestudier efter avslutad gymnasieskola. Kvinnliga gymnasieelever planerar också i betydligt högre utsträckning än manliga att läsa vidare på universitet eller högskola.

## 1.2 Mål och syfte

Med bakgrunden ovan är syftet i denna rapport att undersöka trender samt se hur viljan att studera vidare efter avslutad gymnasieutbildning beror på olika bakomliggande faktorer.

Målet med detta arbete är att få följande frågeställningar besvarade:

- Vilka faktorer påverkar mest respektive minst till att personer väljer att studera vidare efter gymnasiet?
- Är effekten av faktorerna lika för alla högskoleförberedande program?
- Är effekten av faktorerna lika mellan båda könen, män och kvinnor?

I detta arbete kommer vi fokusera på elever som studerat ett högskoleförberedande program och som valt att studera vidare inom ett år efter avslutade gymnasiestudier.

Genom att skapa en prediktionsmodell kommer vi att försöka få en uppfattning för att kunna besvara frågorna ovan. Vidare kommer vi att ta hänsyn till år 2020 som var ett pandemiår och undersöka dess påverkan på om elever valt att studera vidare.

För detta arbete har vi fullständig data för åren 2014–2020 att arbeta med, varför vill vi då göra en statistisk analys? Samt vad kommer vi få ut från vår analys då vi redan är erhållna all data?

Även fast vi har fullständig data kan vi tänka oss att man kan ha singlat slant och fått detta utfall, där vi tänker oss att de individer som gick ut skolan ett visst år var endast ett stickprov på alla möjliga individer som skulle kunna ha gått ut skolan just de året. Det är rimligt att tänka sig att det finns en bakomliggande sannolikhetsfördelning som vi inte känner till och detta råkar vara ett utfall från den sannolikhetsfördelningen, där sannolikhetsfördelningen påverkas av flera olika faktorer. Därav kommer vi i detta arbete undersöka datamaterialet och se vilka faktorer det är som gör att elever väljer att studera vidare ett år efter avslutad gymnasial utbildning.

# 2 Teori

## 2.1 Generaliserad linjär modell

Den generaliserade linjära modellen (GLM) används för att beskriva sambandet mellan en responsvariabel samt en eller flera förklarande variabler.

Den generaliserade linjära modellen är en utvidgning av linjär regression till fall då responsvariablernas inte har en normalfördelning.

De tre komponenter den generaliserade linjära modellen består av är följande (Agresti 2013, kapitel 4.1.1):

- En slumpmässig komponent.
- En systematisk komponent.
- En länkfunktion.

### 2.1.1 Slumpmässig komponent

Den slumpmässiga komponenten identifierar responsvariabeln $Y$ samt dess sannolikhetsfördelning och består av oberoende observationer $(y_1, \ldots, y_N)$ som kommer från den naturliga exponentialfamiljen. En fördelningen som tillhör exponentialfamiljen har en täthets- eller sannolikhetsfunktion, i det kontinuerliga fallet av typ

$$f(y_i; \theta_i) = a(\theta_i)b(y_i)\exp\left[y_i Q(\theta_i)\right],$$

där parametern $\theta_i$ varierar mellan observationerna $i = 1, \ldots, N$ beroende på deras värden på de förklarande variablerna, och vidare kallas $Q(\theta_i)$ för den naturliga parametern.

### 2.1.2 Systematisk komponent

Låt $x_{ij}$ vara värdet av den förklarande variabeln $j$ för observation $i$, där $j = 0, 1, \ldots, p-1$. Detta ger oss att den systematiska komponenten relaterar en vektor $\eta = (\eta_1, \eta_2, \ldots, \eta_N)$ som beskriver sammanhanget mellan den slumpmässiga komponenten samt de förklarande variablerna genom den linjära modellen

$$\eta_i = \sum_{j}^{p-1} \beta_j x_{ij}. \tag{1}$$

Fallet $x_{i0} = 1$ för alla $i$ är interceptet $\beta_0 = \alpha$. Denna linjärkombination kallas för linjär prediktion.

### 2.1.3 Länkfunktion

Den sista komponenten, länkfunktionen, sammansluter den systematiska och de slumpmässiga komponenterna. Länkfunktionen sammansluter $E\left[Y_i \mid X = x\right] = \mu_i$ tillsammans med $\eta_i$ där $\eta_i = g(\mu_i)$. Länkfunktionen, $g$ är differentierbar samt monoton.

Detta ger oss att länkfunktionen $g$ sammansluter $\mu_i$ till de förklarande variablerna genom formeln

$$g(\mu_i) = \sum_{j}^{p-1} \beta_j x_{ij}.$$

Länkfunktionen, $g(\mu_i) = Q(\theta_i)$ som transformerar medelvärdet för den naturliga parametern benämns för kanonisk-länk (Agresti 2013, kapitel 4.1.2).

I den generaliserade linjära modellen behöver inte responsvariabeln, $Y$ vara binär, utan den kan även vara kontinuerlig. I fallet med den generaliserade linjära modellen vill vi skatta ett betingat väntevärde, $\pi(x) = E[Y|X = x]$ som vid en logistisk regression blir en betingad sannolikhet, $\pi(x) = P(Y = 1 \mid X = x)$.

## 2.2 Odds samt oddskvot

Oddskvoten beskriver sambandet och förhållandet mellan två händelser i två olika grupper. De används för att vi ska kunna få en uppfattning om styrkan i sambandet mellan två händelser.

För en lyckad sannolikhet $\pi = \pi(x) = P(Y = 1 \mid X = x)$, där $0 < \pi < 1$, definieras oddsen enligt följande:

$$\Omega = \frac{\pi}{1 - \pi}.$$

Om $\Omega > 1$ innebär det att en framgång är mer sannolikt än en förlust.

Antag en $2 \times 2$-tabell, det vill säga en tabell som innehåller två rader och två kolumner. Oddsen för framgång istället för förlust är $\Omega_i = \frac{\pi_i}{(1-\pi_i)}$. Förhållandet mellan oddsen för rad ett och rad två ges av

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)},$$

och kallas för Oddskvoten.

Fallet då $\theta > 1$ tyder på att händelsen $Y = 1$ är mer sannolik i rad 1 än i rad 2. För fallet då Fallet då $\theta < 1$ gäller de motsatta. När $\theta = 1$ i denna modell så kan vi dra slutsatsen att tillhörigheten till en viss grupp inte har någon effekt på utfallet av $Y$ (Agresti 2013, kapitel 2.2.4).

## 2.3 Logistisk regression

Antag att $X$ är en prediktorvariabel vilken mäts på en intervallskala som antingen antar heltalsvärden eller reella värden. Låt $Y$ vara en binär responsvariabel med utfallen $Y = 0$ eller $Y = 1$, i dessa fall är logistisk regression en lämplig modell att använda.

Den logistiska regressionsmodellen är följande

$$\pi(x) = P(Y = 1 \mid X = x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}.$$

Typiskt val av länkfunktionen $g$ är inversen till logit, så att den linjära logistiska modellen

$$\text{logit}[\pi(x)] = \log \frac{\pi(x)}{1 - \pi(x)} = \alpha + \beta x, \tag{2}$$

är en linjär funktion av $x$ där $\alpha$ är interceptparametern och $\beta$ är effektparametern. Notera att om $X$, $x$ samt $\beta$ antas vara vektorer i ekvation (2), fås en multivariat modell (Agresti 2013, kapitel 5.3.2).

### 2.3.1 Multipel logistisk regression

Den multipla logistiska regressionsmodellen används för att prediktera den binära kategoriska variabeln som grundar sig på flera förklarande variabler. Den ser ut som följande:

$$\pi(x) = \frac{\exp(\alpha + \sum_{j=1}^{p-1} x_j \beta_j)}{1 + \exp(\alpha + \sum_{j=1}^{p-1} x_j \beta_j)},$$

där (Agresti 2013, kapitel 5.4)

$$\begin{aligned}
p &= \text{antalet parametrar,} \\
\beta &= \text{parametervektor,} \\
\alpha &= \text{intercept,} \\
\beta_j &= \text{effektparametern för oberoende variabler } j = 1, \ldots, p-1, \\
x &= (x_1, \ldots, x_{p-1}) = \text{värden av } p-1 \text{ oberoende variabler.}
\end{aligned}$$

### 2.3.2 Oddskvot i logistisk regression

Antag för en multipel logistisk regressionsmodell att den förklarande variabeln $x_1$ är kategorisk. Låt $x_1 = 0$ i den första kategorin och låt $x_1 = 1$ i den andra

kategorin. Om de resterande förklaringsvariablernas värden antas vara lika så får vi att logaritmen av oddskvoten är

$$\begin{aligned}\text{logit}(\pi(x \mid_{x_1=1}) - \text{logit}((\pi(x \mid_{x_1=0}) =& \alpha + \beta_1 \cdot (1) + \ldots + \beta_{p-1}x_{p-1} \\ & - (\alpha + \beta_1 \cdot (0) + \ldots + \beta_{p-1}x_{p-1}) = \beta_1.\end{aligned}$$

Därav är oddskvoten given som $\exp\{\beta_1\}$ (Agresti 2013, kapitel 5.4.1).

## 2.4 Likelihood

För att skatta parametrarna $\beta_j$ använder sig den generaliserade linjära modellen av maximum likelihood-metoden i den systematiska komponenten (1) i modellen.

Antag att att vi har $n_i$ oberoende observationer som kan uttryckas med $x_i = (x_{i0}, \ldots, x_{i,p-1})$. Låt $Z_i$ vara summan av alla deras responsvariabler. Det gäller då att $Z_i \sim \text{Bin}(n_i, \pi(x_i))$. Om $Z_1, \ldots, Z_I$ är oberoende får vi en oberoende binomial radsampling av en $I \times 2$-kontingenstabell. Detta ger vidare likelihoodfunktionen

$$l(\beta) = \prod_{i=1}^{I} P(z_i \mid x_i) = \prod_{i=1}^{I} \binom{n}{z_i} \pi(x_i)^{z_i}(1 - \pi(x_i))^{n_i - z_i},$$

samt log-likelihoodfunktionen

$$\begin{aligned}L(\beta) &= \log\left[\prod_{i=1}^{I} P(z_i \mid x_i)\right] = \sum_{i=1}^{I} \log\left\{\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right][1-\pi(x_i)]^{n_i}\right\} + \text{konstant} \\ &= \sum_{i=1}^{I}\left\{y_i \sum_{j=0}^{p-1}\beta_j x_{ij} - n_i \log\left[1 + \exp(\sum_{j=0}^{p-1}\beta_j x_{ij})\right]\right\} + \text{konstant}.\end{aligned}$$

Maximum likelihood (ML) skattningen $\hat{\beta}_{ML}$ av $\beta$ får vi fram genom att maximera antingen likelihooden eller log-likelihood funktionen. För att beräkna ML-skattningen, går det att använda sig av score-funktionen som är första ordningens derivata av log-likelihood-funktionen som definieras som

$$S(\beta) = \frac{\partial}{\partial\beta} l(\beta) = \sum_{i=1}^{n} \frac{\partial}{\partial\beta} l_i(\beta) = \sum_{i=1}^{n} S_i(\beta).$$

För att få ut skattningen av $\hat{\beta}_{ML}$ sätt $S(\beta) = 0$ och lös ekvationssystemet med $p$ ekvationer och $p$ okända parametrar (Agresti 2013, kapitel 5.5.1).

## 2.5 Akaikes informationskriterium, AIC

Akaikes informationskriterium, AIC, används för att jämföra samt utvärdera olika modeller. Med hjälp av AIC kan vi granska hur väl en modell passar till

en given datamängd och samtidigt ta hänsyn till antal parametrar som finns i modellen. Dess formel är

$$\text{AIC} = -2\ln(L) + 2p,$$

där $L$ är den maximerade logaritmen av sannolikheten för datan och $p$ är antalet parametrar som finns i modellen. Ett lägre värde på AIC indikerar en bättre modell. Detta gör att AIC är ett viktigt verktyg för att jämföra och välja mellan olika modeller (Agresti 2013, kapitel 6.1.6).

## 2.6 Förklaringsgrad, $R^2$

Den vanligaste anpassningsmåttet relaterat till linjära modeller i allmänhet samt multipla regressionsmodeller är den så kallade förklaringsgarden, $R^2$. Modellens förklaringsgrad definieras som den andel av den totala variationen som modellen "förklarar". Vi kan uttrycka $R^2$ med kvadratsummor så att (Sundberg 2022, kapitel 3.2.2)

$$R^2 = \frac{\text{Kvs(regression)}}{\text{Kvs(total)}} = 1 - \frac{\text{Kvs(residual)}}{\text{Kvs(total)}}.$$

### 2.6.1 McFaddens $R^2$

McFadeens $R^2$ används som ett mått för att se kvaliteten på en statistisk modell. Den används för att få en förståelse för variationen i en beroende variabel med hjälp av en uppsättning oberoende variabler.

För att bedöma modellens passform används denna metod inom logistisk regression. Modellen jämför sannolikheten för en logistisk regressionsmodell med endast sannolikheten för en nollmodell, en modell som endast innehåller en interceptterm.

McFaddens $R^2$ definieras som (McFadden 1974, ss. 105–142)

$$1 - \left(\frac{\text{Log-likelihood för modellen}}{\text{Log-likelihood för baslinjen}}\right).$$

## 2.7 Cohens kappa

Cohens kappa används för att bedöma överenskommelsen mellan två mätningar. Cohens kappa jämför sannolikheten för överensstämmelse, $\sum_a \pi_{aa}$ med det förväntade, $\sum_a \pi_{a+}\pi_{+a}$, där mätningarna är oberoende genom

$$\kappa = \frac{\sum_a \pi_{aa} - \sum_a \pi_{a+}\pi_{+a}}{1 - \sum_a \pi_{a+}\pi_{+a}} = 1 - \frac{1 - \sum_a \pi_{aa}}{1 - \sum_a \pi_{a+}\pi_{+a}}.$$

Nämnaren är lika med täljaren med $\sum_a \pi_{aa}$ ersatt av dess maximalt möjliga värden 1, vilket motsvarar perfekt överensstämmelse.

Kappa-värdet, $\kappa$, kan anta värdet 1 i fallet då mätningarna är överens och värdet 0 i fallet då det inte finns en överenskommelse mellan mätningarna annat än vad som skulle förväntas av en slump. Kappa kan även anta värdet $-1$, vilket indikerar en perfekt motsägelse, det vill säga mätningarna är totalt motsatta (Agresti 2013, kapitel 10.5.4).

## 2.8 Bayesianska informationskriteriet, BIC

Bayesianska informationskriteriet, BIC, är ett mått som används för att kunna bestämma ett modellval. En av de viktigaste funktionerna hos BIC är dess att den bestraffar modeller med fler antal parametrar hårdare än jämfört med AIC. Formeln för BIC ser ut som följande

$$\text{BIC} = -2\log(L) + \log(n)k,$$

där $L$ är maximum likelihood av modellen, $p$ är antalet parametrar i modellen och $N$ är stickprovsstorleken. BIC straffar mer komplexa modeller genom att inkludera en straffterm för antalet parametrar i modellen. Detta straff ökar ju fler parametrar som används i modellen. Den modell med lägst BIC-värde är den modell som föredras, eftersom lägst värde indikerar en bättre passform för modellen (Müller, Scealy och Welsh 2013, kapitel 3.3).

# 3 Databeskrivning

Datamaterialet för detta arbete kommer från Skolverket (2022) och täcker åren 2014–2020. Datan innehåller information om antalet elever som slutfört en gymnasieutbildning med behörighet att studera vidare på universitet eller högskola inom ett år efter avslutad gymnasieexamen.

Antalet elever som vi kommer att ta hänsyn till i detta arbete är uppdelade i sex olika högskoleförberedande program, Ekonomi, Estetiska, Humanistiska, International Baccalaureate, Naturvetenskap, Samhällsvetenskap samt Teknik. Vidare kommer vi att ta hänsyn till vilken nationell bakgrund eleverna tillhör samt vad deras föräldrar har som högsta utbildningsnivå. I tabell 1 visualiseras de variabler som kommer ingå i detta arbete samt deras förkortade beteckningar.

Tabell 1: Variabler samt dess beteckningar i rapporten.

| Variabler och beteckningar | | |
|---|---|---|
| Variabler | Beteckning | Typ |
| Totalt | Antalet gymnasieelever som tagit studenten inom ett program, N | Kontinuerlig |
| År | X: $x_i \in [2014, 2020], i = 14, \ldots, 20$ | Numerisk |
| Kön | M = Man<br>K = Kvinna | Kategorisk |
| Program | EK = Ekonomi<br>ES = Estetiska programmet<br>HU = Humanistiska programmet<br>IB = International baccalaureate<br>NA = Naturvetenskap<br>SA = Samhällsvetenskapsprogrammet<br>TE = Teknik | Kategorisk |
| Föräldrars högsta utbildningsnivå | FG = Förgymnasialutbildning<br>GU = Gymnasialutbildning<br>EGKU = Eftergymnasialutbildning kort utbildning<br>EGLU = Eftergymnasialutbildning lång utbildning | Kategorisk |
| Bakgrund | Utl = Utländsk bakgrund<br>Sv = Svensk bakgrund | Kategorisk |

I tabell 1 definieras totalen som samtliga elever som gått ut gymnasiet med en behörighet till högskola med minst 2500 betygsatta poäng. De program som är inkluderade i tabellen ovan är alla de högskoleförberedande programmen som gymnasieeleverna har studerat. Föräldrars högsta utbildningsnivå är följande:

- Förgymnasial utbildning (genomgången folk- eller grundskola).
- Gymnasial utbildning (genomgången gymnasieutbildning).

- Kort eftergymnasial utbildning (eftergymnasial utbildning med minst 30 högskolepoäng och kortare än tre år).
- Lång eftergymnasial utbildning (minst tre år lång eftergymnasial utbildning).

För kategorin bakgrund finns två delgrupper, elever med svensk bakgrund samt elever med utländsk bakgrund. De elever som är födda i Sverige och har minst en förälder som är svenskfödd klassas som svensk bakgrund och elever som är födda utomland eller födda i Sverige men har att båda föräldrarna är födda utomlands klassas som utländsk bakgrund.

Tabell 2: Mängden elever som examinerats år 2014.

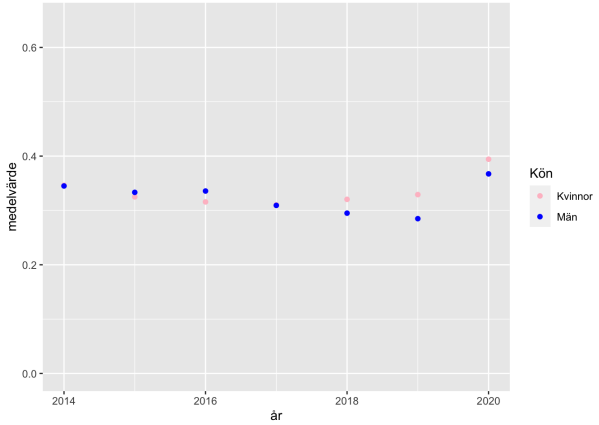
| Program | Kön | Antal i univ. | Antal ej i univ. | Totalt | Andel (%) |
|---|---|---|---|---|---|
| EK | Kvinna | 1278 | 2800 | 4078 | 31.36341 |
| EK | Man | 976 | 2585 | 3561 | 27.43611 |
| ES | Kvinna | 830 | 3717 | 4547 | 18.27579 |
| ES | Man | 365 | 1763 | 2182 | 16.77360 |
| HU | Kvinna | 274 | 406 | 680 | 40.29412 |
| HU | Man | 65 | 90 | 155 | 41.93548 |
| NA | Kvinna | 3055 | 2982 | 6037 | 50.62117 |
| NA | Man | 2971 | 2537 | 5508 | 53.95788 |
| SA | Kvinna | 3081 | 7282 | 10363 | 29.73077 |
| SA | Man | 1283 | 3962 | 5245 | 24.46139 |
| TE | Kvinna | 443 | 661 | 1104 | 40.21739 |
| TE | Man | 2192 | 2670 | 4862 | 45.10490 |
| IB | Kvinna | 130 | 281 | 411 | 31.87348 |
| IB | Man | 79 | 169 | 248 | 31.85484 |

I tabell 2 ser vi data för år 2014över hur stort antal som valt att studera vidare ett år efter de tagit gymnasieexamen och vice versa. Kolumnen *Totalt* visar totalt antal elever som examinerats år 2014 och kolumn *Andel* visar andelen elever av de totala som valt att studera vidare.

## 3.1 Datavisualisering

Med denna studie kommer vi att undersöka vilka faktorer som påverkar gymnasieelever, som studerat ett högskoleförberedande program att ta valet att studera vidare inom ett år efter avslutad gymnasieutbildning mellan åren 2014–2020. Vi kommer att undersöka andelen män och kvinnor, andelen med en svensk eller utländsk bakgrund samt andelen som studerar vidare beroende på vad deras föräldrar har för högsta utbildningsnivå.

Till att börja med räknar vi ut ett medelvärde på andelen elever som väljer att studera vidare ett år efter avslutad gymnasieutbildning. I figur 1 kan vi se medelvärdet för andelen elever som valt att studera vidare ett år efter gymnasiet mellan åren 2014–2020, utan hänsyn till program, föräldrars högsta utbildning samt bakgrund.




Figur 1: Medelvärdet av andel elever som studerar vidare.

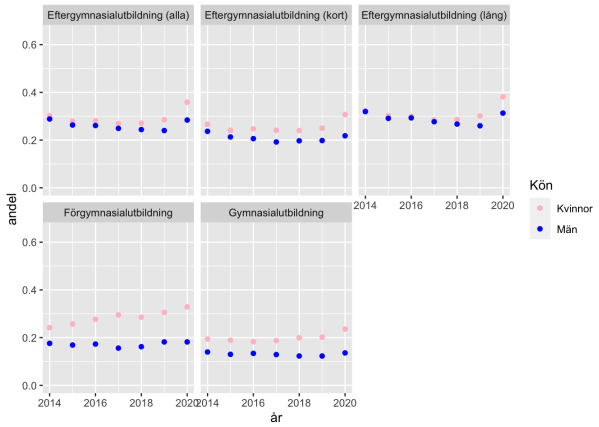
Från figur 1 ser vi att medelvärdet av andelen elever som generellt sett studerar vidare efter avslutad gymnasieutbildning är rätt lika, för båda könen, för åren 2014–2020.

Vidare kan vi se att år 2014 och år 2017 är medelvärdet av andelen elever som studerar vidare rätt lika medan år 2015 och 2016 är det mer manliga elever som dominerar. Efter år 2018 fram till år 2020 sker det en ändring, där medelvärdet av andelen kvinnliga elever dominerar.

För att få en inblick samt en bättre uppfattning i hur andelen är uppdelad mellan åren 2014 till 2020 visualiserar vi datan i nedan kommande plottar.

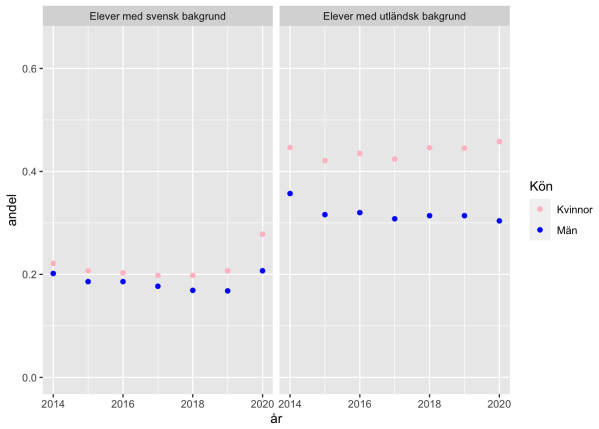
Figur 2: Andelen elever som studerar vidare inom olika program.

Figur 2 ovan visar andelen gymnasieelever som påbörjat studier inom ett år efter gymnasieexamen för de olika högskoleförberedande programmen. Från figuren ser vi att trenden är rätt stabil, generellt ser vi att andelen elever som väljer att studera vidare är hyfsat i jämvikt, vilket är relevant för att kunna dra slutsatsen om att detta är en pålitlig studie. Det vi dock ser är att år 2019 minskar andelen män som studerat vidare efter gymnasiet om de under sina gymnasieår gått det Humanistiska programmet. Totalt sett kan vi se att det sker en ökning för de elever som studerat EK, ES, SA samt kvinnliga elever inom programmet HU. För programmen NA, TE samt IB ser vi att det sker en minskning medan andelen ökar cirka åren 2019–2020.




Figur 3: Andelen elever som väljer att studera vidare beroende på vad deras föräldrar har för högsta utbildningsnivå.

Figur 3 visar andelen gymnasieelever som påbörjat studier inom ett år efter avslutad gymnasieutbildning beroende på vad deras föräldrar har för högsta utbildningsnivå. Från figuren ovan kan vi se att andelen kvinnliga elever som har föräldrar som endast studerat eftergymnasial utbildning ökar för varje år förutom år 2018 då det sjunker en liten del. De elever vars föräldrar som högst studerat en gymnasieutbildning, en kort eftergymnasial utbildning eller en lång eftergymnasial utbildning har varierande andel som studerar vidare. Något utmärkande vi kan se är att år 2020 har andelen kvinnor som studerat vidare ökat markant inom alla typer av utbildningar som ens föräldrar kan ha.




Figur 4: Plottar för Andelen elever som studerar vidare beroende på bakgrund.

I figur 4 ser vi att elever med en utländsk bakgrund har en större andel som studerar vidare för varje år, både för män och kvinnor. Något vi kan lägga märke till är att andelen kvinnliga elever med en utländsk bakgrund som väljer att studera vidare har sjunkit och ökat mellan åren men mellan år 2014–2019 och år 2020 har de ökat. För manliga elever som har en utländsk bakgrund ser vi att det sker en markant minskning efter år 2014 och sedan dess minskar det en del och ökar en del i perioder. Dock är andelen som studerar vidare generellt högre för manliga elever som har en utländsk bakgrund jämfört med de som har en svensk bakgrund. Elever med svensk bakgrund har en mindre andel som studerar vidare och något märkvärdigt är att år 2020 ökar både de kvinnliga och manliga eleverna som studerar vidare markant.

Figur 4 ger även en generell uppfattning om att elever med utländsk bakgrund har en högre andel fortsatta studier efter gymnasiet jämfört med elever som har en svensk bakgrund. Detta kan bero på olika faktorer, såsom att elever med utländsk bakgrund kan ha en högre motivation och engagemang för att lyckas i skolan och för att kunna anpassa sig till det nya samhället. Å andra sidan är det också viktigt att ta hänsyn till att elever med utländsk bakgrund kanske behöver kämpa hårdare för att få en behörighet till högskolan. Det kan finnas

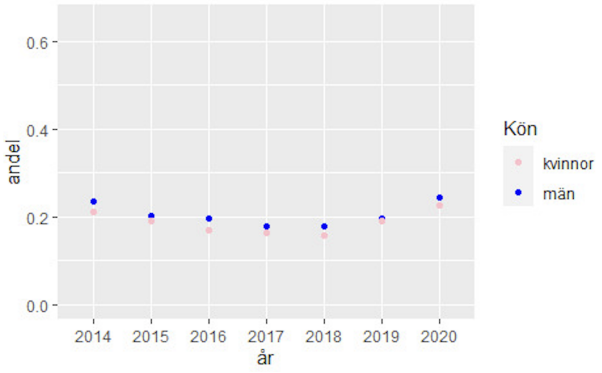
språkliga eller kulturella barriärer som kan påverka deras prestationer i skolan, och de kan möta diskriminering eller fördomar på grund av sin bakgrund. Därför är det viktigt att ta hänsyn till de olika faktorer som kan påverka elevernas prestationer och studieresultat.

För att få en förståelse av arbetslöshetens omfattning under de här åren, presenteras nedan en tabell med procentuell arbetslöshet som hämtats från Statistiska centralbyrån (2023). Tabellen ger en översikt över arbetslöshetens omfattning i procent och hur den utvecklats under den undersökta tidsperioden.

Tabell 3: Procentuell arbetslöshet för män och kvinnor under åren 2014–2020.

| Kön | Ålder | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 |
|---|---|---|---|---|---|---|---|---|
| Män | 15–24 | 23,5 | 20,4 | 19,7 | 17,9 | 17,8 | 19,7 | 24,4 |
| Kvinnor | 15–24 | 21,1 | 19,0 | 16,9 | 16,5 | 15,9 | 19,1 | 22,5 |
| Män | 25–34 | 8,6 | 8,3 | 7,9 | 7,6 | 7,2 | 6,3 | 9,5 |
| Kvinnor | 25–34 | 8,2 | 8,1 | 7,3 | 7,5 | 7,4 | 7,8 | 9,4 |
| Män | 35–44 | 5,4 | 5,1 | 5,0 | 5,2 | 4,5 | 5,0 | 6,1 |
| Kvinnor | 35–44 | 5,7 | 5,6 | 5,4 | 5,1 | 5,3 | 5,2 | 7,1 |
| Män | 45–54 | 5,1 | 4,9 | 4,9 | 4,5 | 4,4 | 4,7 | 4,9 |
| Kvinnor | 45–54 | 4,8 | 4,5 | 4,2 | 3,9 | 3,9 | 4,7 | 5,8 |
| Män | 55–64 | 6,5 | 6,4 | 6,4 | 6,1 | 5,3 | 5,4 | 6,9 |
| Kvinnor | 55–64 | 4,7 | 4,7 | 4,5 | 4,5 | 3,9 | 4,3 | 5,2 |

Tabell 3 ger en översikt över den procentuella arbetslösheten under den undersökta perioden för detta arbete. Vi ser att det exempelvis sker en minskning av arbetslösheten för män mellan åldrarna 15-24 för varje kommande år fram till år 2019. Dessutom visar tabellen att år 2020 skedde en procentuell ökning av arbetslösheten för både män och kvinnor, jämfört med föregående år, inom alla åldersgrupper.




Figur 5: Procentuell arbetslöshet för män och kvinnor mellan ålder 15–24.

Eftersom vi i detta arbete endast ska undersöka de elever som studerat vidare ett år efter avslutad gymnasieutbildning väljer vi att plotta den procentuella arbetslösheten för män och kvinnor mellan åldrarna 15-24 i figur 5.

# 4 Modellering och analys

Från figur 2 ser vi att andelen som studerat vidare ökar inom alla gymnasieprogram år 2020 bortsett från de manliga elever som studerat International Baccalaureate där det sker en minskning. I figur 3 kan vi exempelvis se att andelen elever som väljer att studera vidare ökar markant för de elever som har föräldrar som studerat en lång eftergymnasial utbildning.

Eftersom det generellt sett sker en ökning år 2020 antar vi detta som ett outlierår. Därav kommer vi nedan att testa modellen för åren mellan 2014–2019, där vi inte tar hänsyn till outlieråret samt testa modellen genom att ta med outlieråret 2020 som en extra förklarande variabel.

Det datamaterial som används för detta arbete kommer från Skolverkets databas men innehåller dessvärre inte data om elever på individbaserat nivå och det finns ingen förbindelse mellan variablerna kön, bakgrund och deras föräldrars utbildning. Exempelvis finns det ingen data för en kvinnlig elev som studerat ekonomiprogrammet med en viss bakgrund och vad hennes föräldrar studerat. Utan med den datan som används kan vi endast se förbindelse mellan kön, år samt program, bakgrund eller föräldrars utbildning.

Eftersom vi inte har datamaterial uppdelat på alla kategorier samtidigt, har vi valt att dela upp datan i tre olika grupper som vi sedan tillämpar en logistisk regressionsmodell på. De grupper som vi delat upp datan i följer nedan:

Grupp 1:

$$\begin{cases} \text{År} \\ \text{Kön} \\ \text{Program} \end{cases}$$

Grupp 2:

$$\begin{cases} \text{År} \\ \text{Kön} \\ \text{Föräldrars utbildning} \end{cases}$$

Grupp 3:

$$\begin{cases} \text{År} \\ \text{Kön} \\ \text{Nationell Bakgrund} \end{cases}$$

## 4.1 Logistisk regression

För att få en uppfattning om vilka faktorer som påverkar elevers val att studera vidare efter en avslutad gymnasieutbildning sätter vi upp en logistisk regressionsmodell som genererar sannolikheten för en individ att studera vidare beroende på vilken grupp hen tillhör.

För att tillämpa en logistisk regressionsmodell behöver vi inte känna till $Y_i$ för varje individ $i$, utan det räcker med att vi känner till varje nivåkombination av elever.

En elev som avslutat en gymnasial utbildning väljer antingen att fortsätta studera vidare eller inte, därav har vi att $Y$ är ett binärt utfall där $Y = 1$ om en elev valt att studera vidare och $Y = 0$ om inte. Således är $Y \sim \mathrm{Be}(\pi(x))$ givet att $X = x$ med sannolikhet $\pi(x) = P(Y = 1 \mid X = x)$ att en elev studerar vidare.

Datamaterialet som används i detta arbete innehåller information om summan av Bernoullifördelade variabler $Y_i$ där vi antar att varje individ gör ett oberoende val. Låt $n_{ijk}$ vara antalet individer i kategorin $x = (i, j, k)$, där $i$ står för program, $j$ för kön och $k$ för år. Låt vidare $Z$ vara summan av dessa elever inom en viss kategori, vilket ger oss att

$$Z \sim \mathrm{Bin}(n_{ijk}, \pi(i, j, k)).$$

En logistisk regressionsmodell är en typ av generaliserad linjär modell som används för att modellera och analysera en binomialt fördelad responsvariabel, med hjälp av en länkfunktion som kallas logit.

Typiskt val av länkfunktionen $g$ ger i vårt fall modellen:

$$\mathrm{logit}(\pi(x)) = \alpha + \sum_{k=1}^{7} \beta_k \cdot \mathbb{1}_{\{x_1=k\}} + \gamma \cdot x_2 + \delta \cdot x_3. \tag{3}$$

där vi har vektorn $x = (x_1, x_2, x_3)$ där plats $x_1$ representerar program som antar värden $1, 2, \ldots, 7$, $x_2$ representerar kön som antar värden $0, 1$, för man eller kvinna samt $x_3$ representerar år.

Vidare antar vi ett linjärt samband i modellen (3) för år, medan program och kön indikerar ett kategoriskt samband där vi får en effekt beroende på vilken kategori vi befinner oss i.

## 4.2 Programvara

I detta arbete har vi använt oss av programmeringsspråket R, där paket som exempelvis tidyverse använts. Tidyverse är en samling av olika paket som används för att underlätta dataanalys samt dess visualisering. Paketet innehåller exempelvis dplyr samt ggplot2, dplyr som används för att manipulera data, ggplot2 som används för att skapa en grafisk visualisering av datan.

Sedan har R använts i detta arbete för att både rensa, strukturera samt sammanfoga datamaterialet som kommit till användning.

För att tillämpa en logistisk regressionsmodell krävs vissa antagande. Dessa är att responsvariabeln är binär, observationerna är oberoende, ingen multikollinearitet, det ska inte finnas extrema outliers, att det finns ett linjärt samband mellan den förklarande variabeln och responsvariabeln och att datamängden är tillräckligt stor.

I datamaterialet som används har vi en binär operation, antingen studerar en elev vidare efter avslutad gymnasieutbildning eller inte där varje elev gör ett oberoende val samt att datamaterialet är tillräckligt stort då det sträcker sig över flera tusen elever mellan åren 2014–2020.

Vidare använder vi oss av funktionen gml() för att tillämpa en logistisk regressionsmodell i R. Syftet med denna funktion är att anpassa en modell som förutsäger sannolikheten för att en binär operation ska ske, i vårt fall är det för en individ att studera vidare eller inte efter en avslutad gymnasieutbildning.

För den logistiska regressionsmodellen har vi använt oss av en binomialfördelning som responsvariabeln (att studera vidare eller inte), samt använt oss av den logiska funktionen *logit* för att skapa en linjär modell av förklarande variabler och förutsäga sannolikheten att studera vidare.

## 4.3 Modellframtagning

Från figurer 2,3 samt 4 ser vi att det generellt sker en ökning av andelen elever som studerar vidare år 2020. Det sker en ökning mellan män och kvinnor inom de olika högskoleförberedande programmen, en ökning bland de elever som har en svensk eller utländsk bakgrund samt en ökning bland elever vars föräldrar har olika nivåer av högsta utbildning. Dessutom ser vi även från tabell 3 att den procentuella arbetslösheten har ökat under år 2020.

År 2020 sticker andelen ut och avviker lite mer jämfört med resterande år, detta kan bero på effekterna av ett pandemiår som exempelvis att oddsen för jobb under det året varit sämre, se tabell 3, därav antar vi detta som ett outlierår.

Vid tillämpning av en logistisk regression på en given datamängd får vi en passande kurva till datan samt en predikterad sannolikhet. Till att börja med tillämpar vi en logistisk regression på samtliga år i datan utan att ha hänsyn till något specifikt år. Detta ger oss då en positiv trend, se figurer 12, 13 samt 14 i appendixet. Vidare från tabeller 14, 15 samt 16 i appendixet, kan vi se koefficienterna för respektive grupp då en logistisk regression tillämpas för samtliga år.

I detta arbete antar vi att det är samma trend oavsett kön och program, kön och nationell bakgrund samt kön och föräldrars högsta utbildningsnivå, vilket inte alltid är nödvändigt. Detta eftersom vi antagit att $\text{logit}[\pi(x)]$ modelleras med exempelvis en koefficient per program, kön och ett linjärt samband för år. Detta skulle vi kunna testa genom att lägga till en samspelsvariabel i model-

len, dessvärre har vi inte gjort det. Genom att ta hänsyn till en samspelsterm skulle vi kunna skapa en bättre uppfattning om samband mellan de förklarande variablerna och det binära utfallet som vi försöker förutsäga. I vårt fall, ifall vi skulle ha med en samspelsterm mellan exempelvis kön och år skulle de kunna ge oss en modell med olika lutningar för män och kvinnor på det linjära sambandet över tid, vilket skulle göra det möjligt att testa om skillnaden i lutning är signifikanta.

För att se ifall 2020 är ett år som sticker ut som vi antagit och för att se hur stor effekt det året har testar vi först modellen för åren mellan 2014–2019 och sedan testar vi modellen genom att ta med år 2020 som en till förklarande variabel.

### 4.3.1 Modell för åren 2014–2019

Från figurer 1, 2, 3 och 4 ser vi att år 2020 ökade andelen elever som valt att studera vidare ett år efter avslutad gymnasieutbildning mer jämfört med åren innan. Från figur 3 ser vi även att år 2020 ökade arbetslösheten för ungdomar mellan åldern 15–24. Arbetslösheten som ökade under år 2020 kan därav vara en avgörande faktor för elevers beslut att studera vidare eller inte.

Därav börjar vi med att anta året 2020 som ett outlierår och skapar en logistisk regressionsmodell för endast åren 2014–2019 utan att ta hänsyn till pandemiåret, 2020.

Från tabellerna som följer nedan kan vi för varje grupp se värdet på skattningarna, standardfelen samt dess $P$-värden.

Tabell 4: Tillämpad logistisk regression på grupp 1 mellan åren 2014–2019.

| Koefficienter | Skattning | Std.fel | $Z$-värde | $P$-värde |
|---|---|---|---|---|
| Intercept | 77.781816 | 4.735704 | 16.425 | 2e-16 *** |
| ES | -0.545038 | 0.017301 | -31.503 | 2e-16 *** |
| HU | 0.650160 | 0.033476 | 19.421 | 2e-16 *** |
| IB | 0.289784 | 0.032830 | 8.827 | 2e-16 *** |
| NA | 1.012792 | 0.012275 | 82.509 | 2e-16 *** |
| SA | 0.017257 | 0.012318 | 1.401 | 0.161 |
| TE | 0.671964 | 0.014409 | 46.635 | 2e-16 *** |
| Män | -0.116880 | 0.008581 | -13.621 | 2e-16 *** |
| År | -0.039086 | 0.002348 | -16.645 | 2e-16 *** |

Från tabell 4 kan vi lägga märke till att $P$-värdet för det samhällsvetenskapliga programmet är relativt högt jämfört med resterande koefficienter, men å andra sidan ser vi även att skattningarna för samhällsvetenskapliga programmet är väldigt långt. Detta betyder att ekonomiprogrammet, vilket i detta fall är vår baslinje, är ungefär densamma som samhällsvetenskapliga programmet och att det är en väldigt liten skillnad på andelen mellan dessa. Observerar vi istället på kolumnen *standardfel* för samhällsvetenskapliga programmet är den i linje

med resterande, därav råkar det endast vara en väldigt liten skillnad mellan samhällsvetenskapliga programmet och ekonomiprogrammet.

Tabell 5: Tillämpad logistisk regression på grupp 2 mellan åren 2014–2019.

| Koefficienter | Skattning | Std.fel | $Z$-värde | $P$-värde |
|---|---|---|---|---|
| Intercept | 45.673469 | 4.279265 | 10.67 | 2e-16 *** |
| Eftergymnasial Lång | 0.337001 | 0.009666 | 34.86 | 2e-16 *** |
| Förgymnasial utbildning | -0.008108 | 0.021333 | -0.38 | 0.704 |
| Gymnasial utbildning | -0.426778 | 0.010374 | -41.14 | 2e-16 *** |
| Män | -0.243552 | 0.007309 | -33.32 | 2e-16 *** |
| År | -0.023198 | 0.002122 | -10.93 | 2e-16 *** |

Tabell 6: Tillämpad logistisk regression på grupp 3 mellan åren 2014–2019.

| Koefficienter | Skattning | Std.fel | $Z$-värde | $P$-värde |
|---|---|---|---|---|
| Intercept | 49.365049 | 4.268242 | 11.57 | 2e-16 *** |
| Elever med UB | 0.941582 | 0.008482 | 111.00 | 2e-16 *** |
| Män | -0.238741 | 0.007282 | -32.78 | 2e-16 *** |
| År | -0.025132 | 0.002117 | -11.87 | 2e-16 *** |

Från tabeller 5 och 6 ovan ser vi skattningarna, standardfelet samt $P$-värdet för grupp två och tre. I detta fall ser vi att alla koefficienter förutom förgymnasial utbildning, har väldigt låga $P$-värden samt är statistiskt signifikanta. Att dessa skattningar för koefficienterna är statistiskt signifikanta innebär att sannolikheten för att resultatet ska beror på slumpen är väldigt låg.

Från tabell 5 ser vi att koefficienten förgymnasial utbildning har ett relativt högt $P$-värde, men vi ser även att skattningen för den koefficienten är väldigt låg. Detta motsvarar att koefficienten och vår baslinje, eftergymnasial utbildning, ungefär är densamma och att det endast finns en väldigt liten skillnad på andelen mellan dessa.

I en logistisk regressionsmodell som förutsäger sannolikheten för att en händelse kommer att inträffa eller inte inträffa, definieras oddskvoten som $\exp\{\beta_1\}$.

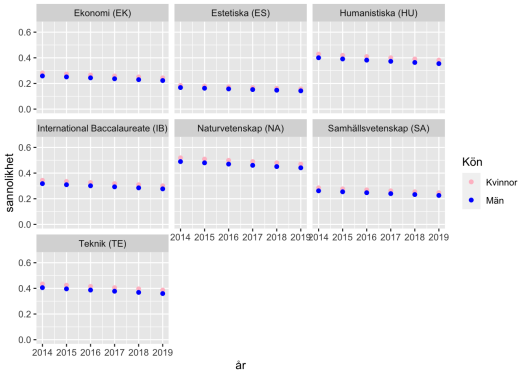
Skattningarna för koefficienterna, $\beta$, är ett mått på hur mycket den predikterande variabeln påverkar logaritmen av oddsen för det binära utfallet. Tag exempelvis skattningen för koefficienten eftergymnasial utbildning (lång), som är 0.337001, vilket säger oss att den predikterade variabeln påverkar logaritmen av oddsen med 0.337001 för utfallet att studera vidare efter avslutad gymnasie studier.

Vidare ser vi att skattningarna för elever med utländsk bakgrund jämfört med vår baslinje, svensk bakgrund, är relativt hög i tabell 6, 0.941582. Även detta

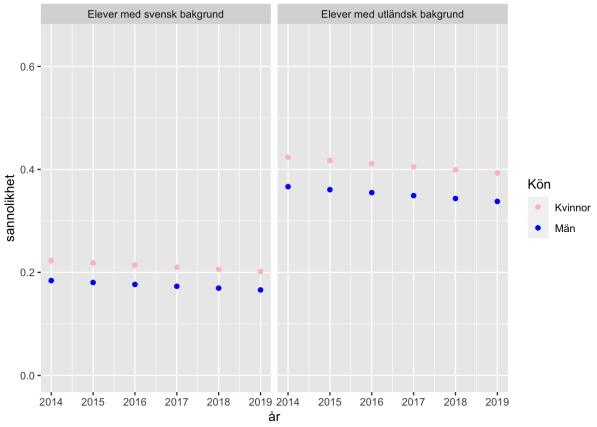
säger oss att den predikterade variabeln påverkar logaritmen av oddsen med 0.941582 för utfallet att studera vidare.

När vi tillämpar en logistisk regressionsmodell är huvudsyftet att lära modellen att förutsäga sannolikheten för en given observation som tillhör en viss klass. Detta eftersom vi ska kunna använda modellen för att göra korrekta förutsägelser på nya observationer.

Därav använder vi oss av den logistiska funktionen för att omvandla logaritmen av oddsen till sannolikheten för det binära utfallet, vilket visualiseras i plottarna som följer nedan för att få en uppfattning om hur sannolikheten ser ut genom åren.




Figur 6: Sannolikheten att gymnasieelever väljer att studera vidare beroende på deras gymnasieprogram.




Figur 7: Sannolikheten att gymnasieelever väljer att studera vidare beroende på deras nationella bakgrund.

Figur 8: Sannolikheten att gymnasieelever väljer att studera vidare beroende på vad deras föräldrar har för utbildning.

Vi kan från figurer 6,7 och 8 se att det är en strikt negativ lutande trend inom alla tre grupper. Vi kan se att sannolikheten att studera vidare minskar för varje kommande år inom alla högskoleförberedande program, inom både elever med svensk och utländsk bakgrund samt för alla elever vars föräldrar har olika höga utbildningsnivåer.

Genom att tillämpa en logistisk regression får vi en passande kurva till datamaterialet. Detta visar vad vi skulle kunna förvänta oss ifall utvecklingen skulle fortsätta på detta vis. I detta fall får vi en negativ lutande trend, vilket skulle kunna ge uppfattningen om sannolikheten att studera vidare förmodligen kommer sjunka för nästkommande år.

För att se hur fallet kommer att bli för nästkommande år 2020 väljer vi att göra om en modell som tar hänsyn till huruvida det var ett pandemiår eller inte.

### 4.3.2 Pandemiår

För att se ifall den negativa trenden fortsätter även år 2020, vilket var ett pandemiår, tar vi hänsyn till det året genom att skapa en modell som har en extra förklarande variabel som testar hur vidare det var ett pandemiår eller inte. Detta eftersom vi vill skapa en uppfattning om hur trenden samt effekten av ett pandemiår kan se ut.

Tabell 7: Värden efter tillämpad logistisk regression på grupp 1 pandemiåret som en extra förklarande variabel.

| Koefficienter | Skattning | Std.fel | $Z$-värde | $P$-värde |
|---|---|---|---|---|
| Intercept | 76.611316 | 4.728730 | 16.201 | 2e-16 *** |
| ES | -0.518285 | 0.015749 | -32.909 | 2e-16 *** |
| HU | 0.669623 | 0.031323 | 21.378 | 2e-16 *** |
| IB | 0.259450 | 0.030217 | 8.586 | 2e-16 *** |
| NA | 1.010417 | 0.011118 | 90.881 | 2e-16 *** |
| SA | 0.036062 | 0.011133 | 3.239 | 0.0012 ** |
| TE | 0.656293 | 0.013055 | 50.272 | 2e-16 *** |
| Män | -0.146262 | 0.007835 | -18.667 | 2e-16 *** |
| År | -0.038501 | 0.002345 | -16.419 | 2e-16 *** |
| Pandemi | 0.394740 | 0.012823 | 30.784 | 2e-16 *** |

Tabell 8: Värden efter tillämpad logistisk regression på grupp 3 pandemiåret som en extra förklarande variabel.

| Koefficienter | Skattning | Std.fel | $Z$-värde | $P$-värde |
|---|---|---|---|---|
| Intercept | 45.951513 | 4.281212 | 10.733 | 2e-16 *** |
| Eftergymnasial Lång | 0.348740 | 0.008859 | 39.367 | 2e-16 *** |
| Förgymnasial utbildning | -0.012278 | 0.019448 | -0.631 | 0.528 |
| Gymnasial utbildning | -0.429317 | 0.009566 | -44.879 | 2e-16 *** |
| Män | -0.277111 | 0.006681 | -41.476 | 2e-16 *** |
| År | -0.023331 | 0.002123 | -10.989 | 2e-16 *** |
| Pandemi | 0.300244 | 0.011646 | 25.780 | 2e-16 *** |

Tabell 9: Värden efter tillämpad logistisk regression på grupp 3 pandemiåret som en extra förklarande variabel.

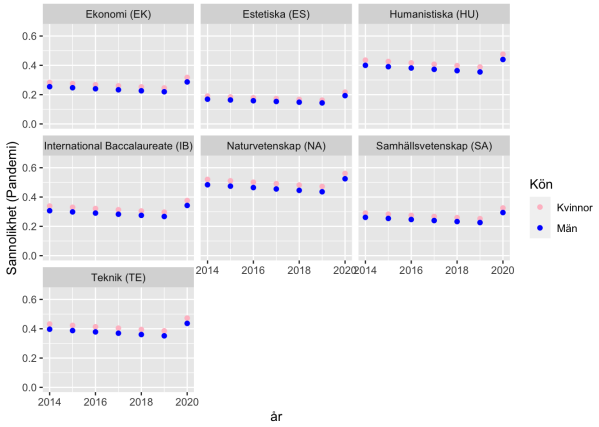
| Koefficienter | Skattning | Std.fel | $Z$-värde | $P$-värde |
|---|---|---|---|---|
| Intercept | 47.696145 | 4.261210 | 11.19 | 2e-16 *** |
| Elever med UB | 0.890580 | 0.007705 | 115.58 | 2e-16 *** |
| Män | -0.276504 | 0.006638 | -41.65 | 2e-16 *** |
| År | -0.024289 | 0.002113 | -11.49 | 2e-16 *** |
| Pandemi | 0.299141 | 0.011559 | 25.88 | 2e-16 *** |

Från tabeller 7, 8 och 9 kan vi se att effekten av att ha ett pandemiår har ett bidrag på cirka 0.30-0.40 för varje kategori, vilket säger att den predikterade variabeln påverkar logaritmen av oddsen genom att öka med cirka 0.30-0.40 för utfallet att studera vidare under ett år då en pandemi pågår.

Vidare ser vi att alla $P$-värden inom alla tre kategorier är statistiskt signifikanta förutom koefficienten förgymnasial utbildning. Även i detta fall har den koefficienten en liten skattning och det säger att den i princip densamma som vår baslinje, eftergymnasial utbildning (kort).

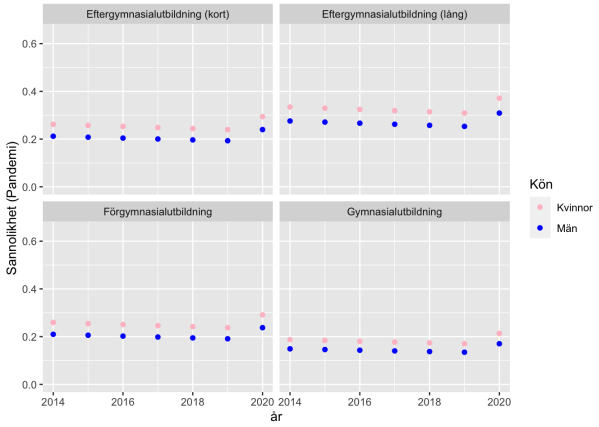
Inom den första kategorin, i tabell 7, ser vi att koefficienten naturvetenskapliga programmet har högst effekt. Inom den andra kategorin, i tabell 8, ser vi att koefficienten eftergymnasial utbildning (lång) har högst effekt samt att de elever som har en utländsk bakgrund, i tabell 9 ger högst effekt inom den tredje kategorin.

Vidare omvandlar vi modellens output till sannolikheter för att kunna göra en bedömning på ifall sannolikheten att studera vidare kommer att sjunka eller stiga under ett pandemiår. Dessa visualiseras i plottarna nedan.



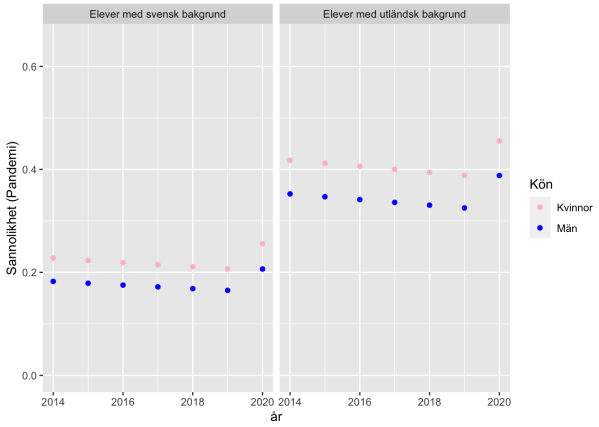

Figur 9: Sannolikheten att gymnasieelever väljer att studera vidare beroende på gymnasieprogram, med pandemiår.

Från figur 9 kan vi lägga märke till att det mellan åren 2014–2019 var en negativ trend, sannolikheten att studera vidare inom alla program minskar för både män och kvinnor. Vidare då vi tar hänsyn till huruvida det var ett pandemiår, kan vi från figur 9 se att andelen kvinnliga och manliga elever som studerat vidare ökat markant inom alla högskoleförberedande program.




Figur 10: Sannolikheten att gymnasieelever väljer att studera vidare beroende på dess nationella bakgrund, med pandemiår.

Figur 10 ovan visualiserar sannolikheten att studera vidare efter gymnasiet för både män och kvinnor beroende på deras föräldrars högsta utbildningsnivåer. Vi kan tydligt se att det för varje år inom varje utbildning finns en minskande trend ända fram till år 2019 medan år 2020 ökar sannolikheten för både män och kvinnor att studera vidare.




Figur 11: Sannolikheten att gymnasieelever väljer att studera vidare beroende på vad deras föräldrar har för utbildning, med pandemiår.

Även för kategori tre ser vi från figur 11 att det är en negativ lutande trend för den predicerade sannolikheten som modellen ger. Sannolikheten att studera vidare minskar fram till år 2020 där det sedan ökar markant.

Genom att tillämpa en logistisk regressionsmodell och undersöka trenden försöker vi förutsäga sannolikheten för framtida observationer. Därav ifall 2020 inte skulle vara ett speciellt år med en pandemi som slog ut skulle vi förmodligen förutspå att även år 2020 skulle det ske en minskande trend i andelen som studerar vidare. Nu blev det inte så, utan effekten av pandemin blev exempelvis att flera blev arbetslösa vilket kan ha resulterat i att allt fler valde att börja studera efter en avslutad gymnasieutbildning.

## 4.4 Mcfaddens $R^2$-värde

För att testa hur väl modellerna som använts i detta arbete passar datamaterialet har vi tagit fram Mcfaddens $R^2$-värden för modellen som tillämpas för alla tre modeller. Ett Mcfaddens $R^2$-värden som är mellan 0.2 och 0.4 anses ofta indikera en god modellanpassning, medan värden under 0.2 anses indikera en dålig passning och värden över 0.4 anses indikera en mycket bra passning (Allison 2014, ss. 1–3) . Eftersom vi är intresserade av att få en uppfattning om hur trenden såg ut innan och efter pandemin, tar vi fram $R^2$-värdet för alla tre modeller som testats i detta arbete.

Tabell 10: $R^2$-värden för samtliga modeller.

| Modeller | $R^2$ (2014–2020) | $R^2$ (2014–2019) | $R^2$ (pandemiåret) |
|---|---|---|---|
| Gymnasiepro | 0.9006793 | 0.9545433 | 0.9467466 |
| Föräldrarsutb | 0.8716412 | 0.9155759 | 0.9194882 |
| Bakgrund | 0.8996145 | 0.9587224 | 0.9405494 |

Från tabell 10 ovan ser vi att $R^2$-värdet för test av alla modeller inom alla kategorier är relativt höga, väldigt nära 1, vilket betyder att vår modell har en bra passform till det datamaterial som använts. Jämför vi modellen som observerar åren 2014–2020 med modellen som tar med pandemiåret som en förklarande variabel ser vi att modellen som tar med pandemiåret som en förklarande variabel har större $R^2$-värde. När vi lägger till pandemiåret i modellen passar den bättre eftersom vi får en extra variabel samt ett extra år som modellen tar hänsyn till, vilket ger oss en bättre uppfattning. Generellt sett ser vi att då modellen endast testas på åren mellan 2014–2019 genererar det bäst $R^2$ värden, vilket indikerar en mycket bra passform på datan enligt McFadden $R^2$-värden.

## 4.5 Cohens kappa

För att få en uppfattning om överensstämmelsen mellan två mätningar tittar vi på modellernas olika kappa-värden. Ett kappa-värde fås genom att vi jämför det faktiska med en slump och är användbart eftersom det tar hänsyn till chansöverensstämmelse. Kappa-värdet är nära 1 i fallet då mätningarna är överens och nära 0 då mätningarna inte alls stämmer överens.

Tabell 11: $\kappa$-värden för samtliga modeller.

| Grupp | $\kappa$ (2014–2020) | $\kappa$ (2014–2019) | $\kappa$ (pandemiåret) |
|---|---|---|---|
| Gymnasiepro | 0.94680 | 0.95584 | 0.97294 |
| Föräldrarsutb | 0.97248 | 0.98052 | 0.98127 |
| Bakgrund | 0.97399 | 0.98466 | 0.98104 |

I tabell 11 ser vi att de olika grupperna inom de olika modellerna får ett relativt högt kappa-värde, alla över 0.95. Detta säger oss att modellerna stämmer överens med mätningarna och att modeller passar bra in jämfört med vad den predikterar.

## 4.6 AIC

För att kunna bestämma kvaliteten på modellerna som används i detta arbete tittar vi på modellernas AIC-värde som finns nedan i tabell 12. Dock för att kunna jämföra AIC-värden, måste vi ha samma datamaterial med lika många observationer, det har vi inte i modellen som testas mellan åren 2014–2019. Även fast vi har med AIC värden för modellen mellan åren 2014–2019 i tabell 12, och ser att de har bättre värden, kan vi endast jämföra modellen mellan 2014–2020 med modellen som har med pandemiåret som en förklarande variabel.

AIC används främst för att kunna jämföra olika modeller och se vilken av de som beskriver data på bästa möjliga vis. AIC ger en indikation på hur bra en statistisk modell passar till datamaterialet samt hur komplex modellen är. En modell med lägre AIC-värde anses ha bättre beskrivning förmåga än en modell med högre AIC-värde (Akaike 1974, ss. 716–721).

Tabell 12: AIC-värden för samtliga modeller.

| Modeller | AIC (2014–2020) | AIC (2014–2019) | AIC (pandemiåret) |
|---|---|---|---|
| Gymnasiepro | 2827.3 | 1454.4 | 1883.4 |
| Föräldrarsutb | 2283.1 | 1299 | 1623.3 |
| Bakgrund | 1917.4 | 796.37 | 1252.3 |

I tabell 12 ovan kan vi se att modellen som tar med pandemiår som en förklarande variabel jämfört med modellen mellan åren 2014–2020 som inte gör det har bättre AIC värde.

## 4.7 BIC

Vidare undersöker vi BIC-värden för de olika modellerna som kan ses i tabell 13 nedan. BIC ger en balans mellan modellens passform och dess komplexitet och det hjälper till att undvika överanpassning av modeller till datamaterialet.

Tabell 13: BIC-värden för samtliga modeller.

| Modeller | BIC (2014–2020) | BIC (2014–2019) | BIC (pandemiåret) |
|---|---|---|---|
| Gymnasiepro | 2850.552 | 1476.282 | 1909.229 |
| Föräldrarsutb | 2295.281 | 1310.184 | 1637.506 |
| Bakgrund | 1922.741 | 801.0785 | 1258.94 |

BIC tar hänsyn till log-likelihood, antalet parametrar i modellen och är en modellvalsmetod med ändamål att hitta den bästa balansen mellan modellens anpassning till data och dess komplexitet. Från ovan, tabell 13, ser vi att modellen som testas mellan år 2014–2019 genererar det bästa BIC-värdet jämfört med resterande modeller. Dock för att jämföra BIC-värden bör modellerna ha samma antal observationer detta eftersom BIC-värdet kan påverkas ifall modellerna man jämför har olika observationer. Exempelvis skulle en modell med lägre observationer ha ett högre BIC-värde eftersom de har mindre data att anpassa sig till, vilket skulle kunna resultera i högre osäkerhet.

Jämför vi istället modellen som tar med pandemiåret som en extra förklarande variabel med den som inte gör det ser vi att modellen som tar hänsyn huruvida det var ett pandemiår genererar ett bättre BIC värde, vilket enligt BIC genererar den bästa modellen.

# 5 Diskussion och slutsats

I detta arbete har vi kommit fram till att det mellan åren 2014–2019 sker en negativ lutande trend i sannolikheten att studera vidare efter avslutad gymnasial utbildning. Vidare har vi kommit fram till att pandemiåret, 2020, var ett år där andelen som studerade vidare ett år efter avslutad gymnasial utbildning ökat. Elever, inom båda könen, med både svensk samt utländsk bakgrund och oavsett vad deras föräldrar haft för högsta utbildning.

Genom tillämpning av en logistisk regressionsmodell har vi kommit fram till att innan pandemin slog ut var det en negativ trend, andelen studenter som valde att studera vidare efter avslutad gymnasial utbildning minskade för varje nästkommande år medan pandemin medförde en förändring.

Även fast regressionsmodellen visar en negativ trend för åren 2014–2019, ser vi i figur 6 att sannolikheten att fortsätta studera ett år efter att ha avslutat en gymnasial utbildning är högre för både män och kvinnor som har studerat ett av programmen NA, TE samt HU, jämfört med resterande högskoleförberedande program. Detta tyder på att en av faktorerna som kan påverka elevernas val att fortsätta studera är att de har valt ett mer krävande program på gymnasiet.

I vidare analys, från figur 9 kan vi se att den högsta andelen elever som studerar vidare efter avslutad gymnasial utbildning förekommer inom naturvetenskapliga programmet sedan följt av teknik samt humanistiska programmet. Exempelvis under år 2020 var sannolikheten att studera vidare cirka 0.5-0.6 för elever som studerat naturvetenskapliga programmet. Följt av teknik och humanistiska programmet som hade en sannolikhet att studera vidare på cirka 0.4-0.5 under år 2020.

Sannolikheten att elever som studerat naturvetenskapliga programmet väljer att studera vidare är generellt sett högre jämfört med resterande högskoleförberedande program. Vilket även kan stärkas med statistik från Berg och Kossack (2020) som skriver:

> Av dem som examinerades från högskoleförberedande program på gymnasiet 2015/16, var det elever på naturvetenskapliga program som i högst grad hade påbörjat eftergymnasiala utbildningar inom tre läsår.

Vidare från figurer 7 och 8, även fast det är en negativ trend, kan vi se att elever med utländsk bakgrund och elever vars föräldrar har genomgått en längre eftergymnasial utbildning har en tendens att ha en högre sannolikhet att fortsätta studera jämfört med elever som har en svensk bakgrund och elever vars föräldrar har en kortare utbildning. Detta tyder på att faktorer som föräldrarnas utbildningsnivå och elevernas nationella bakgrund kan spela en roll för deras beslut att fortsätta sin utbildning.

Vidare från figur 10 kan vi se att elever med utländsk bakgrund tenderar att ha en högre sannolikhet att studera vidare jämfört med elever som har en svensk bakgrund. I figur 10 ser vi exempelvis att en kvinnlig elev som har en utländsk

bakgrund har år 2016 en sannolikhet på cirka 0.4 medan en kvinnlig elev som har en svensk bakgrund samma år har en sannolikhet på cirka 0.23.

En möjlig förklaring till varför elever med utländsk bakgrund dominerar när det gäller att fortsätta studera efter gymnasiet är för att de inte tar sin utbildning för givet. Elever med bakgrund från länder där utbildningsmöjligheterna är begränsade, eller där vissa grupper inte har rätt till utbildning, kan ha en större uppskattning för en utbildning. De kan ha upplevt svårigheter eller begränsningar när det gäller tillgång till utbildning och därför värderar de möjligheten att studera högre än elever som inte har upplevt sådana begränsningar. Hansson (2018) skriver:

> Personer med utländsk bakgrund har länge varit underrepresenterade bland studenter på högskolor och universitet i Sverige, men de senaste åren har statistiken vänt. För många är utbildning något värdefullt, dyrt och ouppnåeligt i hemländerna. Jag kommer från en bakgrund där vi inte tar sådant för givet och därför har en helt annan syn på utbildning, samt tror att de med svensk bakgrund tar demokrati och utbildning för givet.

Att flera med utländsk bakgrund väljer att studera vidare kan även bero på andra faktorer, exempelvis skulle det kunna bero på att flera elever med utländsk bakgrund, mellan åldrarna 15–24 är arbetslösa och därför väljer att studera. Dessutom skulle det kunna finnas en chans att flera elever med utländsk bakgrund har svårare att uppnå en behörighet att studera vidare och därmed inte ingår i vår statistik.

Efter att ha genomfört en ytterligare analys för att undersöka hur ett pandemiår påverkar elevernas beslut att fortsätta studera vidare, kan vi från figurerna 9, 10 och 11 se att sannolikheten att fortsätta studera ökar inom alla program, oavsett om eleverna har svensk eller utländsk bakgrund och oavsett om deras föräldrar har olika nivåer av högsta utbildning. Vilket säger oss att det pandemin medfört har haft en positiv effekt på elevernas beslut att fortsätta sin utbildning, trots de utmaningar från själva pandemin. Det finns en tydlig pandemieffekt, och vi kan förvänta oss att denna inom några år kommer klinga av, och att andelen som studerar vidare inom ett år återgår till nivåer innan pandemin, allteftersom arbetslösheten bland unga fortsätter sjunka.

En möjlig förklaring till varför fler personer väljer att fortsätta studera under pandemiåret är för att antalet arbetslösa har ökat markant under de året, vilket framgår i figur 3. Detta framgår av Kamath, Stener och Lennartsson (2020, s. 1):

> Under 2020 präglades hela det svenska samhället av utbrottet av covid-19, och arbetsmarknaden var inget undantag. Pandemin har gjort ett tydligt avtryck i arbetsmarknadsstatistiken när året summeras, men det har också varit stor variation under årets gång. Under våren försämrades läget på arbetsmarknaden kraftigt och nådde en botten under sommaren. [...] Coronapandemin har förändrat be-

> teenden i samhället och har slagit hårt mot vissa branscher och särskilt de unga på arbetsmarknaden. I genomsnitt var 5 064 000 personer sysselsatta under 2020. Jämfört med 2019 är det en minskning med 67 000 personer, varav 51 000 var ungdomar i åldern 15-24 år.

Den modell utan hänsyn till pandemiåret som passas till datamaterialet för perioden 2014–2020 ger en positiv trend över tid för andelen som väljer studera vidare, medan den modell med en indikator för pandemiår samt den modell för perioden 2014–2019 båda ger en negativ trend. En visuell inspektion av datamaterialet ger en tydlig bild av en negativ trend fram till 2020 inom alla grupper, och detta bekräftas i analysen i och med att modellen för 2014–2020 utan indikator för pandemiåret har sämre AIC- och BIC-värden.

Slutligtvis för att kunna göra framtida prognoser och se trender skulle det därav vara mest lämpligt att använda sig av ett datamaterial som innehåller åren mellan 2014–2019, utan en påverkan av ett pandemiår, eller använda fullständig data mellan åren 2014–2020 som vi gjort fast med en extra förklarande variabel. Elever som har studerat NA, TE eller HU på gymnasiet har högre sannolikhet att fortsätta sin utbildning. Dessutom har elever med utländsk bakgrund och de vars föräldrar har en lång eftergymnasial utbildning också en högre sannolikhet att fortsätta studera.

Vidare kan vi från resultatet se att pandemin i sig inte haft en påverkan på elevers beslut att studera, utan det som kan ha påverkat är vad pandemin medfört, som exempelvis att flera människor blivit arbetslösa samt haft minskad chans till anställning. Under pandemin har det skett en ökning i sannolikhet att fortsätta studera inom alla program, för både svenska och utländska bakgrunder, samt för elever vars föräldrar har olika nivåer av högsta utbildning.

En viktig synvinkel att ta hänsyn till är att datamaterialet som används i detta arbete kommer från Skolverket som inte är individbaserat. Detta innebär att datamaterialet inte innehåller förbindelse mellan elevernas kön, nationell bakgrund, föräldrars högsta utbildningsnivå eller högskoleförberedande program som de studerat. Om datamaterialet istället varit individbaserad skulle vi kunna använda en logistisk regressionsanalys för att jämföra flera koefficienter och få en mer utförlig förklaring till vilka faktorer som påverkar elevers beslut att studera vidare efter gymnasieutbildningen. Exempelvis skulle vi kunna undersöka skillnaden mellan kvinnliga elever med svensk bakgrund som studerat ekonomiprogrammet med en kvinnlig elev med utländskbakgrund som studerat ekonomiprogrammet, och dra slutsatser om vilken elev som är mer benägen att studera vidare efter gymnasiet.

I efterhand, efter avslutad studie skulle det ha varit intressant om modellen som används haft något mått på arbetslöshet som en förklarande variabel. Möjligen hade den då förklarat stor del av pandemieffekten. Det skulle i så fall också göra det möjligt att prediktera när pandemieffekten sjunkit till före pandeminivå, genom att gissa sig till när arbetslösheten sjunkit till den nivå den var innan pandemin.

Utöver det har vi i detta arbete inte testat samspel mellan variablerna för varje grupp. Exempelvis skulle vi för grupp ett kunna testa samspel mellan år och kön eller mellan år och program. Genom att ta hänsyn till en samspelsterm skulle vi kunna skapa en bättre uppfattning om sambandet mellan de förklarande variablerna och det binära utfallet som vi försöker förutsäga. I vårt fall skulle samspelstermerna vara användbara för att få en djupare förståelse av hur de olika förklarande variablerna interagerar och påverkar sannolikheten för en händelse att inträffa. I vårt fall skulle vi kunna få olika linjära samband över tid beroende på kategori. Dessutom skulle det vara möjligt att testa om skillnaderna mellan lutningen i varje kategori (till exempel man eller kvinna) är signifikant, eller om den modell vi satt upp utan detta samspel är lika väl passande.

# Referenser


Agresti, A. (2013) *Cathegorical Data Analysis* (3:e uppl.) Hoboken, NJ: Wiley.

Akaike, H. (1974) "A New Look at the Statistical Model Identification". *IEEE Transactions on Automatic Control*, vol. 19:6, ss. 716–723.

Allison, P. D. (2014) "Measures of fit for logistic regression". *Statistical Horizons LLC and the University of Pennsylvania*.

Berg, D. och Kossack, P. (2020) "Färre män läser vidare efter gymnasieskolan". *Statistiska centralbyrån*. Hämtad 2023-04-03 från `https://www.scb.se/hitta-statistik/statistik-efter-amne/utbildning-och-forskning/befolkningens-utbildning/overgang-gymnasieskola-eftergymnasiala-utbildningar/pong/statistiknyhet/overgang-gymnasieskola--eftergymnasial-utbildning-examinerade-fran-gymnasieskolan/`.

Hansson, F. (2018) "Studie: Fler med utländsk bakgrund studerar vidare". *Dagens Nyheter*. Hämtad 2023-04-03 från `https://www.dn.se/ekonomi/jobb-karriar/studie-fler-med-utlandsk-bakgrund-studerar-vidare/`.

Inkinen, M. och Amnéus, I. (2016) "Kvinnor och män i högskolan". *Universitetskanslersämbetet*. Hämtad 2023-04-15 från `https://gamla.uka.se/download/18.5d85793915901d205f9a852/1487841854015/rapport-2016-10-14-kvinnor-och-man-i-hogskolan.pdf`.

Jacobsson, G., et al. (2008) "Kvinnor och män i högskolan". *Högskoleverket*. Hämtad 2023-04-04 från `https://docplayer.se/2651-Rapport-2008-20-r-kvinnor-och-man-i-hogskolan.html`.

Kamath, M., Stener, L. och Lennartsson, F. (2020) "Arbetsmarknadssituationen för hela befolkningen 15-74 år, AKU 2020". *Statistiska centralbyrån*. Hämtad 2023-05-05 från `https://www.scb.se/contentassets/e9b07f6a123d4da0a2a92d8d77d5f15f/am0401_2020a01_sm_am12sm2101.pdf`.

McFadden, D. (1974) "Conditional Logit Analysis of Qualitative Choice Behavior". *Frontiers in Econometrics*. New York, NY: Academic Press, ss. 105–142.

Müller, S., Scealy, J. L. och Welsh, A. H. (2013) "Model Selection in Linear Mixed Models". *Statistal Science*, vol. 28:2, ss. 135–167.

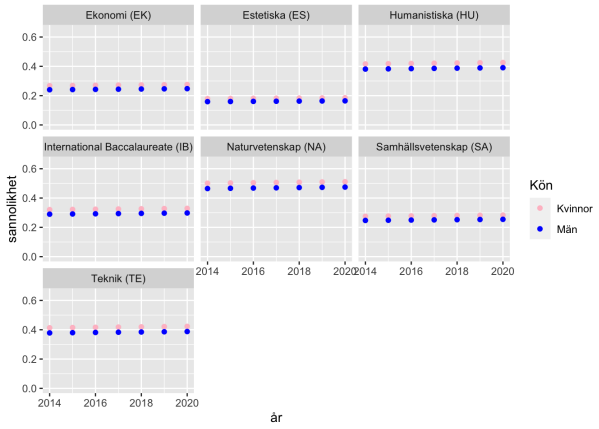
Skolverket (2022) *Sök statistik*. Hämtad 2023-02-02 från `https://www.skolverket.se/skolutveckling/statistik/sok-statistik-om-forskola-skola-och-vuxenutbildning?sok=SokC&verkform=Gymnasieskolan&omrade=Betyg%20och%20studieresultat&lasar=2014/15&run=1`.

Sundberg, R. (2022) *Linjära stokastiska modeller*. Stockholm: Stockholms universitet.

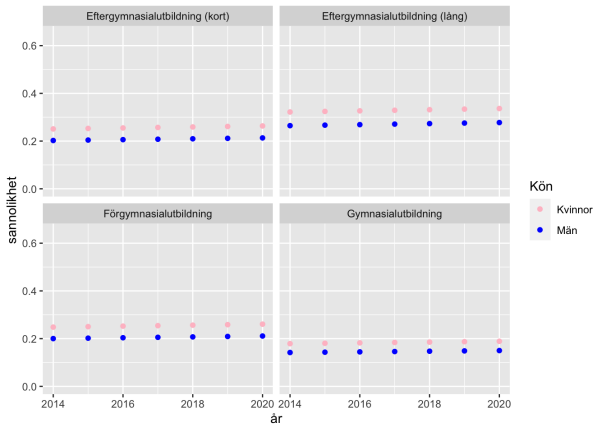
Statistiska centralbyrån (2023) *Befolkningen 15-74 år (AKU) efter kön, ålder och arbetskraftstillhörighet. År 2001 - 2022*. Hämtad 2023-05-02 från `https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__AM__AM0401__AM0401A/NAKUBefolkning2Ar/`.

Söderman, G. (2022) "Nybörjare i högskolan 2021/22". *Universitetskanslersämbetet*. Hämtad 2023-05-02 från https://www.uka.se/download/18.2a2ae249185be5f310fcf72/%201674219508472/Nybrjare%20i%20hgskolan%202021-22.pdf.

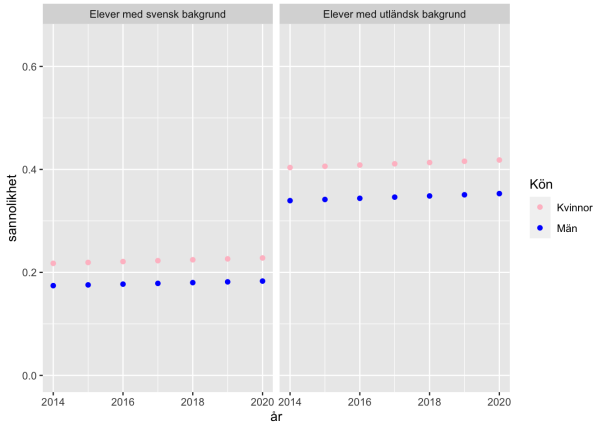
# Appendix




Figur 12: Program.




Figur 13: Föräldrars utbildning.




Figur 14: Bakgrund.

Tabell 14: Värden efter tillämpad logistisk regression på grupp 1 pandemiåret som en extra förklarande variabel.

| Koefficienter | Skattning | Std.fel | $Z$-värde | $P$-värde |
|---|---|---|---|---|
| Intercept | -14.393894 | 3.682369 | -3.909 | 9.27e-05 *** |
| ES | -0.514488 | 0.015731 | -32.705 | 2e-16 *** |
| HU | 0.666796 | 0.031282 | 21.315 | 2e-16 *** |
| IB | 0.254815 | 0.030175 | 8.444 | 2e-16 *** |
| NA | 1.010006 | 0.011101 | 90.987 | 2e-16 *** |
| SA | 0.037558 | 0.011117 | 3.378 | 0.000729 *** |
| TE | 0.654758 | 0.013035 | 50.231 | 2e-16 *** |
| Män | -0.145803 | 0.007825 | -18.634 | 2e-16 *** |
| År | 0.006648 | 0.001825 | 3.642 | 0.000271 *** |

Tabell 15: Värden efter tillämpad logistisk regression på grupp 3 pandemiåret som en extra förklarande variabel.

| Koefficienter | Skattning | Std.fel | $Z$-värde | $P$-värde |
|---|---|---|---|---|
| Intercept | -23.135101 | 3.330728 | -6.946 | 3.76e-12 *** |
| Eftergymnasial Lång | 0.349002 | 0.008853 | 39.424 | 2e-16 *** |
| Förgymnasial utbildning | -0.013912 | 0.019435 | -0.716 | 0.474 |
| Gymnasialutbildning | -0.428851 | 0.009561 | -44.856 | 2e-16 *** |
| Män | -0.276979 | 0.006677 | -41.485 | 2e-16 *** |
| År | 0.010944 | 0.001651 | 6.627 | 3.42e-11 *** |

Tabell 16: Tillämpad logistisk regression på grupp 3 mellan åren 2014–2019.

| Koefficienter | Skattning | Std.fel | $Z$-värde | $P$-värde |
|---|---|---|---|---|
| Intercept | -21.548130 | 3.308705 | -6.513 | 7.39e-11 *** |
| Elever med UB | 0.889458 | 0.007698 | 115.545 | 2e-16 *** |
| Män | -0.275817 | 0.006633 | -41.580 | 2e-16 *** |
| År | 0.010064 | 0.001640 | 6.135 | 8.52e-10 *** |