LUNDS UNIVERSITET

Ungas läsning

En logistisk regressionsanalys av hur socioekonomi och födelseland påverkar läsningen bland unga

Statistiska institutionen
Lunds universitet
Kandidatuppsats i statistik
Nivå 61–90 15 hp
Handledare: Yvette Baurne

Vanessa Sevedag
HT2023

Abstract

Previous reports and studies have found that gender, socioeconomic status, migration background and education are important factors in explaining reading amongst young people. A previous report from the Swedish Agency for Youth and Civil Society indicate that there are differences in reading depending on gender and birth country using confidence intervals. However, the report indicated that living in an area with higher socioeconomic status are not associated with more reading among young people in Sweden. In this thesis I examine the same data using binary logistic regression.

The dependent variable is measured by young people in Sweden who read books at least every week, using the results from the National Youth Survey. The results show that reading is dependent on gender, birth country (Sweden or abroad), parents' education, socioeconomic status of living area as well as library visits. Very high socioeconomic status of living area is, however, negatively associated with reading compared to low socioeconomic status of living area. A possible explanation is the disproportionate level of young people with parents with higher education amongst respondents.

Nyckelord: *läsning, unga, socioekonomi, födelseland, föräldrars utbildningsnivå*

Innehåll

1 Inledning

Målet för Sveriges ungdomspolitik är att alla unga, 13–25 år, ska ha goda levnadsvillkor, makt att forma sina liv och inflytande över samhällsutvecklingen (Regeringens proposition 2013/14:191). Myndigheten för ungdoms- och civilsamhällesfrågor, MUCF, har i uppdrag att följa upp ungdomspolitiken genom ett antal indikatorer som belyser ungas levnadsvillkor som helhet. Uppföljningen av ungdomspolitiken används som underlag i utformning av insatser för att främja goda levnadsvillkor för alla unga. För att kunna utforma träffsäkra insatser är det av intresse om det finns skillnader i levnadsvillkor och om förutsättningar för goda levnadsvillkor skiljer sig åt mellan unga beroende på exempelvis kön, socioekonomi, eller mellan unga som är inrikes eller utrikes födda. En återkommande utmaning i analysen av vad som påverkar levnadsvillkoren är att flera faktorer ofta hänger samman, som exempelvis födelseland och socioekonomi.

Ett prioriterat område för ungdomspolitiken är att alla unga ska ha en meningsfull fritid (Regeringens skrivelse 2020/21:105). En av indikatorerna för ungas fritid är andelen unga som läser varje vecka. Indikatorn baseras på nationella ungdomsenkäten som skickas ut av MUCF var tredje år. Utöver att vara en meningsfull fritidsaktivitet kan läsning ses som en färdighet som har betydelse för ungas övriga levnadsvillkor. I ett betänkande från Läsdelegationen sammanfattar de att läsning är viktigt för unga för

> *”…att kunna navigera i världen, att skaffa sig nödvändig kunskap och information, men också om att känna läsningens glädje och njutning. Läsning är av central betydelse för såväl privatliv och arbetsliv som utbildning och samhällsliv. Både i och utanför skolan spelar läsförmågan en viktig roll för barns och ungas kunskapsutveckling och möjlighet att såväl nu som i framtiden kunna ta del av och påverka det samhälle de lever i.”* (SOU 2018:57)

Att så många unga som möjligt läser, och att läsningen är jämt fördelad mellan olika grupper av unga kan alltså ses som betydelsefullt för att nå målsättningen om att alla unga ska ha goda levnadsvillkor.

En tidigare analys av ungas läsning visar dock att det finns skillnader i ungas läsning. Tjejer läser i högre utsträckning än killar och att unga utrikes födda läser i högre uträckning än inrikes födda (MUCF, 2023). I analysen undersöktes även om det fanns skillnader i läsning mellan unga i områden med olika socioekonomiska förutsättningar och skillnader mellan unga beroende på ålder. Det fanns ingen skillnad mellan unga i olika åldrar. Det är sedan tidigare välkänt att det finns skillnader i levnadsvillkor kopplat till socioekonomi. Flera skillnader mellan unga i områden med olika socioekonomiska förutsättningar framkom också i samma undersökning för andra variabler. Det var därför oväntat att resultatet inte visade på att unga från områden med socioekonomiska utmaningar läste i lägre utsträckning än andra unga. Istället visade resultatet att unga i områden med socioekonomiska utmaningar läste i högre utsträckning än unga med bättre förutsättningar. En möjlig förklaring till att det inte finns skillnader i läsning är att andelen utrikes födda är högre i områden med socioekonomiska utmaningar än i områden med goda socioekonomiska förutsättningar.

Eftersom den tidigare analysen baseras konfidensintervall, med i huvudsak endimensionella analyser. Anledningen till att analyserna i huvudsak presenteras endimensionellt är på grund av att konfidensintervallen då flera variabler tas hänsyn till ofta blir så pass breda att i princip inga skillnader är signifikanta. Mot bakgrund av detta är det möjligt att sammansättningen av

utrikes födda i de olika områdestyperna osynliggör skillnader kopplat till socioekonomi. En fördjupad analys av variablerna skulle kunna bidra till att öka förståelsen för variationen i läsning bland unga, genom att undersöka flera variabler i samma modell.

1.1 Syfte och frågeställning

Mot denna bakgrund vill jag i denna uppsats undersöka hur födelseland och socioekonomisk områdestyp tillsammans förklarar ungas läsning. Jag vill även undersöka hur ungas biblioteksbesök samvarierar med läsning.

- Vilken metod är bäst lämpad för att undersöka hur födelseland och socioekonomisk områdestyp påverkar läsning hos unga?
- Vilka slutsatser är möjliga att dra utifrån en fördjupad analys som tar hänsyn till flera variabler?
- Bidrar en sådan analys till en bättre förståelse för skillnader i läsning mellan olika grupper av unga?

Uppsatsens resultat kan användas som vägledning för hur framtida analyser på bästa sätt kan utformas för att bättre besvara frågeställningar där flera bakgrundsvariabler påverkar, eller kan påverka, den beroende variabeln.

1.2 Disposition

Uppsatsen är indelad i 7 kapitel samt en referenslista och en bilaga. I *kapitel 1* ges en beskrivning av bakgrunden till uppsatsens ämne, syfte och frågeställningar. *Kapitel 2* innehåller beskrivning av tidigare forskning om ungas läsning samt tidigare studier som använt sig av, för uppsatsen, relevanta metoder. I *kapitel 3* presenteras den data som kommer att användas. Även förekomst och hantering av bortfall diskuteras. *Kapitel 4* innehåller beskrivning och diskussion om val och utvärdering av metod samt en diskussion om alternativa metodval. I *kapitel 5* ”Deskriptiv statistik” presenteras relevant deskriptiv statistik om datamaterialet som kommer att användas i analysen. Därefter utvecklas och utvärderas analysmodellen i *kapitel 6.* I kapitlet diskuteras även eventuella problem med den valda modellen. I *kapitel 7* ”Avslutande diskussion” sammanfattas uppsatsens resultat. Reflektioner över modellens och datamaterialets svagheter görs.

2 Bakgrund

I detta kapitel ges en kort beskrivning av befintlig kunskap om ungas läsning. Ett urval tidigare studier om läsning och kön, socioekonomi och migrationsbakgrund presenteras kortfattat. Ett antal studier som använt relevanta metoder presenteras kortfattat. Detta kapitel är inte en uttömmande beskrivning av befintlig forskning, utan syftar till att göra nedslag som är relevanta för ämnet.

2.1 Befintlig kunskap om barns och ungas läsning och bibliotekstillgång

Tidigare forskning och rapporter om ungas läsning fokuserar framför allt på barns och ungas läsförståelse, läsning och modersmål samt skillnader mellan killars och tjejers läsning. Det finns även studier som undersöker bibliotekstillgång och socioekonomiska faktorer. Ett annat område som undersökts är barn och ungas läsnings påverkan på bland annat kreativitet och förmåga att relatera till andra. Många studier på området är kvalitativa, men det förekommer även kvantitativa studier.

Det finns begränsat med forskning som undersöker andelen unga som läser på fritiden. Tidigare undersökningar om hur ofta unga läser, eller andelen som läser görs främst av myndigheter (se exempelvis Statens Medieråd, 2023; MUCF, 2023; MUCF, 2021). Studierna visar på att andelen unga som läser har minskat under en längre tid, och att tjejer läser i högre utsträckning än killar. En undersökning från Statens medieråd visar dock att andelen unga som läser har ökat något under 2022 (Statens Medieråd, 2023). Att myndigheter undersöker hur ofta unga läser kan ses mot bakgrund av att de är politiskt styrda organisationer, där det finns ett politiskt intresse av att unga läser. Den statliga utredningen *Barns och ungas läsning – ett ansvar för hela samhället* (SOU 2018:57) beskriver läsning som en grundläggande färdighet som har stor betydelse för många olika delar av ungas liv.

Kön och genus betydelse för barn och ungas läsning har flera studier från olika länder belyst på olika sätt (Se exempelvis Fischer, 2022; Kurnaz & Pursun, 2022; Ee Loh, Sun & Majid, 2020; Lepper, Stang-Rabrig och McElvany, 2022). Ett flertal svenska kandidatuppsatser inom olika ämnen undersöker även detta ämne kvalitativt (Se exempelvis Tiberg & Trulsson, 2021; Skoggren, 2022).

Det finns även ett flertal studier som på olika sätt berör läsning, socioekonomi och migration. Ett par av studierna fokuserar på barn och ungas läsförståelse och har följt barn och unga över tid (Lathouras, Westerveld &Trembath, 2019; Barone, Fouge & Pin, 2021).

Flera av studierna visar på att flera faktorer påverkar barns och ungas läsning samtidigt. En turkisk studie visar på att flera faktorer påverkar läsmotivation bland unga gymnasieelever, såsom kön, betyg och läsning på fritiden. Faktorer såsom mammans yrke, familjens inkomst eller antalet lästa böcker årligen hade däremot ingen påverkan på läsmotivation (Kurnaz & Pursun, 2022). En irländsk studie visar på att kön och familjens sociala klass påverkar barns läsförståelse, och att sambandet mellan ungas läsning, kön och social klass är additiva, där barn från högre social klass och tjejer har bättre läsförståelse. Insatser i skolan och hur ofta föräldrar läser för sina barn bidrog till att förklara skillnader (McGrinnity, 2022).

I en kandidatuppsats i kulturgeografi används socioekonomiska områdestyper för att undersöka bibliotekstillgång i Stockholm. Uppsatsen är särskilt intressant eftersom den använder samma indexvariabel för socioekonomi som kommer att användas i denna uppsats. Slutsatsen i uppsatsen i kulturgeografi är att bibliotekstillgången är något bättre i områden

med socioekonomiska utmaningar och stora utmaningar jämfört med områden med mycket goda socioekonomiska förutsättningar. En möjlig delförklaring till skillnaden kan vara att hustyperna skiljer sig mellan områdestyperna, där områden med socioekonomiska utmaningar till stor del består av flerbostadshus medan områden med (mycket) goda socioekonomiska förutsättningar i större utsträckning består av villor. Detta gör att områden med goda socioekonomiska förutsättningar är mindre tätt befolkade och att avståndet till biblioteket i området blir större, jämfört med om samma antal hushåll bor i flerbostadshus (Forsell, 2023).

2.2 Tidigare kvantitativa studier och metodval

Det finns flera studier som analyserat liknande variabler som kommer att användas i denna uppsats, såsom läsning, föräldrars utbildningsnivå och socioekonomisk områdestyp. Eftersom socioekonomisk områdestyp är en relativt ny indexvariabel är det av intresse att se hur denna har använts i tidigare studier och med vilka resultat. Hur tidigare studier valt att analyser frekvens av läsning är också intressant, eftersom den beroende variabeln är styrande för val av metod.

En brittisk studie har undersökt vad som påverkar om föräldrar läser för sina barn genom logistisk regression. Medan studien analyserar föräldrars läsning för barn, snarare än barn och ungas läsning, är val av metod och eventuella hinder intressant, eftersom den beroende variabeln liknar den beroende variabel som kommer att användas i denna uppsats. Den beroende variabeln bestod av en enkätfråga på 6-gradig skala, där föräldrar besvarade hur ofta del läste för sina barn, från aldrig till varje dag. Den beroende variabeln delades senare in i fyra kategorier, eftersom frekvensen för vissa svarsalternativ var låga. Studien använde sig av en ordinal logistisk regressionsmodell, det vill säga en logistisk regression där den beroende variabeln analyserades med fler än två kategorier. Hänsyn togs till variabelns ordinala karaktär (Fischer, 2022).

En annan studie som undersökt läsfrekvens har studerat danska femteklassares läsning och analyserat samband med bland annat kön, socioekonomi, migrationsbakgrund genom en random effects regressionsmodell. Den beroende variabeln är i den här studien, till skillnad från den brittiska studien, numerisk eftersom läsning är mätt i antal minuter den unga läser via en app. Studien var dock tvungen att avslutas i mars 2020 på grund av att pandemin medförde svårigheter att skilja läsning på fritiden och i skolan. Sju olika modeller undersöktes med kön som utgångspunkt. Ett antal interaktionsvariabler testades, såsom kön och akademisk attityd, kön och läsförståelse samt när läsning skedde. Studien visade att tjejer läser mer än killar utanför skoltid (Smith & Reimer, 2023).

En tidigare studie som har använt sig av socioekonomisk områdestyp är forskarna Abdelzadeh och Lundberg som undersökt unga utifrån bland annat socioekonomiska skillnader och skillnader mellan unga med svensk respektive utländsk bakgrund och föräldrarnas utbildningsbakgrund är *Ungas röst* (Abdelzadeh & Lundberg, 2020) som skrevs på uppdrag av MUCF. Där undersökte de bland annat skillnader i ungas valdeltagande med hjälp av en logistisk regressionsanalys. Socioekonomi analyserades genom variabeln socioekonomisk områdestyp (1–5). Föräldrars utbildningsbakgrund delades in i en sjugradig skala från förgymnasial kortare än 9 år till forskarutbildning. Samtliga faktorer analyserades som dummyvariabler. För de kategorivariabler som hade fler än två kategorier användes en kategori som jämförelsekategori med samtliga andra. För socioekonomi användes områden med sämst socioekonomiska förutsättningar som jämförelsepunkt. Rapporten visar på

skillnader i valdeltagande mellan unga beroende på kön, svensk respektive utländsk bakgrund, föräldrars utbildningsnivå och unga från områden med olika socioekonomiska förutsättningar.

Tidigare forskning och andra studier indikerar på att kön, födelseland och socioekonomiska faktorer samt studierelaterade faktorer är relevanta faktorer för att förklara olika aspekter ungas läsning. Det finns dock inga studier som undersöker kön, socioekonomi, föräldrarnas utbildningsbakgrund och födelseland samtidigt i förhållande till hur ofta unga läser. Tidigare studier som undersökt läsfrekvens har använt sig och olika typer av regressionsanalys beroende på om beroende variabeln är numerisk eller ickenumerisk. För studien som använt en ickenumerisk beroende variabel har antal observationer har påverkat indelningen av beroende variabeln. I de studier som har använt socioekonomisk områdestyp har alla kategorier använts. I studien om valdeltagande har en kategori, unga i områdestyp 1, använts som jämförelsekategori. Att tidigare studier har goda erfarenheter att använda samtliga områdeskategorier indikerar på att detta alternativ bör undersökas i första hand. Hur logistisk regression har använts på en beroende variabel som är mycket lik den som används i denna uppsats gör att ordinal regression bör undersökas som alternativ för val av metod.

3 Data

I detta kapitel ges en närmare beskrivning av den data och variablerna som kommer att användas som underlag. Därutöver beskrivs bortfall och hantering av bortfallet.

3.1 Nationella ungdomsenkäten

Underlaget för uppsatsen är Nationella ungdomsenkäten 2021. Nationella ungdomsenkäten är en enkätundersökning som genomförs var tredje år av MUCF. Enkäten som genomfördes 2021 skickades ut till ett slumpmässigt urval om 12 000 unga 16–25 år, varav cirka 50 procent besvarade enkäten.

Enkäten kunde besvaras via en pappersenkät som skickades ut till de unga eller via webben. Det fanns även möjlighet att besvara enkäten på engelska. En majoritet besvarade enkäten digitalt och endast ett fåtal besvarade enkäten på engelska. 3 påminnelser skickades ut, 2 på sms och 1 per brev.

De som besvarade enkäten fick ett presentkort på 100 kr.

En aspekt som är viktig att påtala är att enkäten genomfördes under covid-19-pandemin.

3.1.1 Beroende variabel

Den beroende variabeln baserar sig på en enkätfråga från nationella ungdomsenkäten: *hur ofta gör du följande på din fritid: läser böcker*. Svarsalternativen utgörs av alternativ på en ordinalskala med fem svarsalternativ, från varje dag till aldrig. Analysen kommer att baseras på en sammanslagning av svarsalternativen. Detta diskuteras närmare i avsnitt 4.1.2.

Tabell 1. Beroende variabeln

Fråga	Underfråga	Svarsalternativ	Indelning för analys
Ungefär hur ofta gör du följande på din fritid:	Läser böcker (inklusive surfplatta)	Varje dag; Varje vecka; Varje månad; Varje år; Aldrig	Varje vecka Mer sällan

3.1.2 Oberoende variabler

De oberoende variabler som kommer att analyseras i denna uppsats är ungas födelseland, kön, föräldrars högsta avslutade utbildning och socioekonomisk områdestyp den unga är bosatt i. Födelseland, kön och föräldrars utbildning är baserade på svar i nationella ungdomsenkäten. Socioekonomisk områdestyp baseras på registerdata som inhämtas om den svarande.

Utöver dessa variabler kommer även besök på bibliotek varje vecka att användas som oberoende variabel. Till skillnad från övriga oberoende variabler är inte möjligt att avgöra riktningen för en eventuell samvariation mellan läsning och biblioteksbesök. Detta hindrar inte att samvariationen undersöks, men det är viktigt att ta hänsyn till detta i de slutsatser som dras.

I tabell 1 och 2 presenteras variablerna närmare.

Tabell 2. Variabler baserade på enkätfrågor.

Fråga	Underfråga	Svarsalternativ	Indelning för analys
Var är du och dina föräldrar födda?	Du själv	Sverige; Övriga Norden; Europa; Utanför Europa; Vet ej/inte aktuellt	Inrikes född Utrikes född
Är du?		Kille; Tjej; Icke-binär; Annan könsidentitet; Osäker	Kille Tjej Ickebinär/annan/osäker
Vilken är din och dina föräldrars högsta avslutade och godkända utbildning?	Förälder 1 Förälder 2	Gick inte ut Grundskolan; Grundskolan (eller motsvarande); Gymnasium (eller motsvarande); Utbildning efter gymnasiet, ej högskola/ universitet; Utbildning vid högskola eller universitet Inte aktuellt/vet inte	Föräldrarnas högsta utbildning: Utbildning vid högskola eller universitet Annan utbildningsnivå
Ungefär hur ofta är du på följande platser?	bibliotek	Varje dag, varje vecka, varje månad, varje år; aldrig	Varje vecka Mer sällan

Tabell 3. Variabler baserade på registerdata.

Registerdata	Indelning
Socioekonomisk områdestyp	Område med mycket goda socioekonomiska förutsättningar Område med goda socioekonomiska förutsättningar Område med blandad socioekonomi Område med socioekonomiska utmaningar Område med stora socioekonomiska utmaningar

Bakgrundsvariabeln för socioekonomisk områdestyp skiljer sig från övriga variabler i underlaget eftersom det dels baseras på registerdata, dels är en indexvariabel. Variabeln baseras på regionala statistikområden, även kallade RegSO-områden, som delar in Sverige i 3363 områden. Dessa områden har kategoriserats i 5 områdestyper efter socioekonomisk sammansättning, där områdestyp 1 har en hög koncentration av personer med sämre socioekonomiska förutsättningar och områdestyp 5 har en hög koncentration av personer med mycket goda socioekonomiska förutsättningar. Områdestyp 3 präglas av en blandning av personer med goda och sämre socioekonomiska förutsättningar, och det är därför ett socioekonomiskt blandat område (Boverket, 2022).

Indexet för att mäta socioekonomisk områdestyp är en sammanvägning av följande variabler:

- Disponibel inkomst per konsumtionsenhet (median)
- Andel personer med låg respektive hög ekonomisk standard
- Andel personer med försörjningsstöd men utan övriga ersättningar
- Andel hemmaboende barn 0–17 år i familjer med låg inkomststandard

- Utbildningsnivå

3.1.3 Bortfall

I Nationella ungdomsenkäten finns ett större bortfall bland killar än tjejer, bland utrikes födda jämfört med inrikes födda. Bortfallet varierar även med ålder. Ett vanligt sätt att hantera denna typ av bortfall är genom att vikta enkäten där svar från de underrepresenterade grupperna viktas upp för att kompensera för bortfallet. Detta förutsätter att fördelningen i befolkningen är känd. Resultatet för nationella ungdomsenkäten kan viktas för juridiskt kön (kvinna eller man), födelseland (inrikes eller utrikes född) och ålder (16–18 år, 19–21 år och 22–25 år). Denna uppsats har som ambition att undersöka variabler som påverkar läsningen, och inte uppskatta läsningen i den unga befolkningen som helhet kommer analyserna göras utan viktning. Eftersom viktningens funktion är att öka betydelsen av svar från underrepresenterade grupper för att skattningarna bättre ska spegla populationens sammansättning bör inte viktningen påverka analysens resultat. Detta hade dock behövt undersökas för att säkerställa att så verkligen är fallet, vilket hade medfört ytterligare analyser som inte utgör det primära syftet med uppsatsen.

I de flesta enkätundersökningar förekommer dels bortfall i form av att vissa individer väljer att inte besvara enkäten, men också i form av bortfall för enskilda variabler. Bortfall utgör ett problem av flera skäl. Dels innebär båda former av bortfall ett mindre antal observationer vilket skapar större osäkerhet i materialet, och därmed ökad varians, vilket bidrar till minskade möjligheter att förkasta nollhypotesen. Bortfallet för de som valt att inte besvara enkäten kan i detta läge inte åtgärdas utan behöver tas hänsyn till i slutsatserna. Bortfall i enskilda variabler kan dock hanteras på olika sätt i ett datamaterial Ett sätt att hantera att det saknas vissa observationer för oberoende variabler kan hanteras antingen genom att utesluta individen ur det analyserade datasetet eller genom imputering.

3.1.3.1 Imputering

Imputering kan göras på flera sätt. En metod är att ersätta den saknade variabeln med ett medelvärde. Detta rekommenderas dock inte eftersom det kan leda till stora fel i estimeringen och en alltför lågt skattad varians. Ett sätt att hantera denna svaghet är multipel imputation, som skapar flera datasätt, där de saknade värdena ersätts med olika värden. Dessa kan analyseras och på så sätt kan skattningar av bortfallet väljas ut. (Hosmer, Lemeshow och Sturdivant, 2013; Osbourne, 2015).

Hosmer, Lemeshow och Sturdivant (2013) rekommenderar att multipel imputering används då det kan antas att bortfallet är slumpmässigt och är medelstort. I datamaterialet finns bortfall gällande områdestyp, födelseland och föräldrars utbildningsnivå. Områdestyperna konstrueras utifrån information RegSO-områden, vilket är registerdata, som utgår från var den unga personen bor. Att information om den ungas boende saknas skulle exempelvis kunna förklaras av att individen för tillfället saknar en folkbokföringsadress. Det finns alltså anledning att tro att informationen inte saknas av slumpen utan att dessa individer avviker från övriga observationer. Det kan alltså vara problematiskt att tillämpa imputation för dessa observationer. Att inte inkludera dessa i analysen kommer troligen få mycket begränsad påverkan på resultatet, då antalet saknade observationer är 14. Födelseland är till skillnad från områdestyp baserat på enkätsvar. Bortfallet kan alltså dels uppstå då individen missar att svara på frågan, dels då individen väljer att inte svara på frågan. Det är av förklarliga skäl svårt att avgöra om det finns ett motiv eller inte till att ett viss individ inte svarat. Eftersom

födelseland är en fråga som av vissa kan uppfattas som känslig går det inte att utesluta att individer väljer att avstå från att svara på frågan. I det fall det rör sig om att individen inte vill svara är bortfallet inte slumpmässigt och individerna avviker i så fall troligen från övriga svarande. Även i detta fall blir det problematiskt att använda imputation. På grund av att bortfallet är litet, 30 observationer, kommer det inte att få någon större påverkan på analysen. Det finns även ett litet bortfall för biblioteksbesök på 14 observationer.

Bortfallet om information om föräldrars utbildning kan delas in i två typer, dels unga som svarat att de inte känner till någon förälders högsta utbildningsnivå, dels unga som inte besvarat frågan. I det första fallet känner vi till varför informationen saknas, eftersom den unga själv uppgett anledningen. Att den unga inte vet om, eller bedömer det som aktuellt, vad deras förälder eller föräldrar har för utbildningsnivå kan bero på olika saker, såsom att föräldern har utbildning från ett annat land med ett utbildningssystem som är svårt att översätta till en svensk kontext, eller att den unga inte har en relation till sin förälder som gör att denne känner till så mycket om förälderns bakgrund. Som framgår av de möjliga förklaringarna kan vi alltså inte anta att deras avsaknad av kunskap är slumpmässig. Att välja att inte besvara just denna fråga, då de besvarat enkäten som helhet kan heller inte antas vara slumpmässigt utan goda skäl. Med andra ord är inte heller detta bortfall lämpligt för imputering.

4 Metod

I detta kapitel beskrivs val av metod samt närmare beskrivning av den metod som kommer att användas. Olika val och möjliga utmaningar diskuteras.

4.1 Val av regressionsmodell

Jag vill undersöka hur socioekonomi och födelseland påverkar ungas läsning samt samvariationen mellan biblioteksbesök och läsning och om detta skiljer sig beroende på socioekonomi och födelseland. Detta gör jag genom att använda mig av en regressionsanalys. Eftersom den beroende variabeln är en diskret ordinal variabel är linjär regression inte ett lämpligt val. Det finns flera andra modeller för regressionsanalysen som skulle var möjliga att använda. Binär, multinominal eller ordinal logistisk regressionsmodell kan vara rimliga alternativ, eftersom de, till skillnad från linjär regression, tar hänsyn till att variabeln inte är numerisk.

En binär logistisk regressionsmodell är anpassad för en analys där det beroende variabeln utgörs av två kategorier, en dummyvariabel. I det här fallet inträffar en viss frekvens av läsning, alternativt gör det inte det. En multinominal logistisk regression fungerar i princip på samma sätt som den binära logistiska regressionsmodellen, med skillnaden att den beroende variabeln antar fler kategorier än två. I en sådan modell behöver ingen sammanslagning av kategorier göras för en variabel som har fler än två kategorier (Hosmer, Lemeshow & Sturdivant, 2013).

Ytterligare ett alternativ är den ordinala modellen. En sådan modell har fördelen att modellen tar hänsyn till att den beroende variabeln (i detta fall läsning) har en rangordning mellan svarsalternativen, där den som läser varje dag läser oftare än den som läser varje vecka och så vidare. En multinominal analys tittar på svarsalternativ som är fler än 2, men tar inte hänsyn till någon rangordning mellan alternativen (Hosmer, Lemeshow & Sturdivant, 2013). Nackdelen med den multinominala modellen är att analysen därmed inte tar hänsyn till mönster kopplat till rangordningen, vilket troligen finns. Eftersom syftet med analysen är att undersöka hur ett antal variabler påverkar ungas läsning (frekvensen), skulle rangordningen vara intressant för slutsatsen.

En utmaning med både den ordinala och multinominala logistiska regressionen är att fler kategorier generellt innebär att antalet observationer inom varje kategori i den beroende variabeln riskerar att bli få, trots att stickprovsstorleken är stor, särskilt då fördelningen mellan kategorierna inte är jämnt fördelat. Exempelvis utgör andelen utrikes födda cirka 11 procent av de som besvarat Nationella ungdomsenkäten och andelen som bor i områden med mycket stora socioekonomiska utmaningar utgör 4.5 procent. I det här fallet är en viktig förklaring till att andelen för dessa delgrupper är liten att andelen är liten i populationen som helhet. Därutöver kan benägenheten hos vissa grupper att svara vara lägre, vilket kan bidra till att vissa delgrupper blir ytterligare mindre. Trots att antalet observationer är stort, över 5000 observationer (efter att felaktiga svar och partiellt bortfall rensats ut), blir antalet unga utrikes födda i områden med mycket goda socioekonomiska förutsättningar 40 individer, vilket innebär att antalet inom de flesta svarskategorierna för läsning blir lägre än 10. Tas hänsyn till ytterligare gruppindelningar som kön blir antalet individer mycket få.

Medan det alltså kan finnas ett värde att analysera läsning med hänsyn till samtliga kategorier, kan det alltså finnas andra faktorer som talar emot att använda en analysmodell där antalet

kategorier bidrar till att delgrupperna blir mycket små. Få observationer bidrar till att skapa osäkerhet i skattningarna och det går att ifrågasätta hur väl dessa fåtal individer kan sägas spegla hela populationen med samma egenskaper, inte minst då syftet med en ordinal regressionsmodell är att fånga mer nyanserade resultat än den binära logistiska regressionsmodellen. Slutsatsen blir därför att binär logistisk regression är den bäst lämpade modellen för syftet med denna uppsats utifrån den data som finns att tillgå.

En ytterligare möjlighet till analys, som avviker från de etablerade metoderna att göra en analys på en icke-numerisk beroende variabel, är att göra en linjär regression. Detta är den metod som idag används på MUCF. Anledningen till att linjär regression inte är en etablerad metod för att analysera en kategorisk beroende variabel är eftersom modellen kan leda till estimerade sannolikheter som ligger utanför 0-1 och inte ger en meningsfull estimering av Pr(Y|X) (Gareth mfl., 2021).

4.2 Binär logistisk regression

Logistisk regression är den vanligast använda regressionsmodellen för att analysera andra variablers påverkan på en diskret variabel. Målet med modellen, är liksom linjär regression, att hitta den bästa modellen för att beskriva förhållandet mellan den beroende variabeln och en uppsättning andra förklarande variabler. Skillnaden mellan metoderna är att logistisk regression är binär i sin karaktär, där den beroende variabelns händelse antingen finns (1) eller inte finns (0), medan en linjär regression undersöker en kontinuerlig, numerär beroende variabel och hur förändringen hänger samman med en eller flera förklarande variabler, vilka kan vara numerära eller kategoriska. För både linjär och logistisk regression antas observationerna vara oberoende, de oberoende variablerna antas vara korrelerade samt att det inte finns outliers som har stor påverkan på modellen[1] (Hosmer, Lemeshow & Sturdivant, 2013).

Den logistiska regressionsmodellen utgår från sannolikheten att y ska inträffa, givet en viss egenskap hos x, E(Y|X)=P(Y|X)= $\pi(x)$. Den beroende variabeln, y, kan därmed bara anta två värden beroende på om den unga läser varje vecka (y=1), eller inte (y=0). $\pi(x)$ utgör sannolikheten att y=1 ska inträffa givet x. E(Y|X) kan därmed enbart anta värden mellan 0 och 1 och följer en binomialfördelning.

$$E(Y|X) = P(Y|X) = \pi(x) = \frac{e^{\beta_0+\beta_1 x_1}}{1 - e^{\beta_0+\beta_1 x_1}}$$

$\pi(x)$ utgör sedan grunden för en logit-transformation för att skapa en linjär modell, vilken får egenskaper som gör att modellen kan utvärderas och analyseras på liknande sätt som en linjär regressionsmodell. Transformationer ger:

$$g(x) = ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 x_1$$

(Hosmer, Lemeshow & Sturdivant, 2013; Gareth mfl., 2021).

[1] På grund av att samtliga variabler är kategoriska går det inte att göra någon ordinarie analys av detta, eftersom extremvärden i vanlig betydelse inte kan förekomma.

4.2.1 Oddskvoter

En oddskvot är ett sätt inom logistisk regression att göra relationen mellan den beroende och de oberoende variablerna begripliga. I linjär regression är det möjligt att utifrån koefficienterna för respektive variabel enkelt förstå hur stor påverkan de oberoende variablerna har på den beroende variabeln. Eftersom den logistiska modellen transformerar något icke-linjärt till en linjär modell behövs ett annat sätt att förstå hur stor påverkan, eller samvariation, de oberoende variablerna har. Oddskvoter är ett sätt att göra detta, genom att jämföra hur ofta det som undersöks inträffar då den oberoende variabeln antar en viss egenskap. En kategori hos den oberoende variabeln utgör jämförelsepunkten (x=0) mot en annan kategori variabeln kan anta (x=1). Oddskvoten beräknas

$$OR = \frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))}.$$

Finns en skillnad mellan $\pi(0)$ och $\pi(1)$ kommer oddskvoten för x=1 att anta ett värde som är mindre eller större än 1 (Hosmer, Lemeshow & Sturdivant, 2013). För att tydliggöra relationen mellan oddskvoten och koefficienten kan oddskvoten också formuleras på följande sätt:

$$OR = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

Om oddskvoten antar ett värde som är större än 1 finns ett positivt samband mellan den beroende och oberoende variabeln. Är oddskvoten mindre än 1 är sambandet negativt. Antar oddskvoten värdet 1 finns inget samband mellan den beroende och oberoende variabeln. (Hosmer, Lemeshow & Sturdivant, 2013).

4.2.2 Binär logistisk regression vid beroende variabel med flera kategorier

I det fall ordinal eller multinominal logistisk regression inte bedöms lämplig återstår är att använda sig av logistisk regression där den beroende variabeln delas in i två kategorier utifrån de flertalet kategorier som finns. Utmaningen handlar framförallt att hitta en indelning som på ett lämpligt sätt fångar det som ska mätas. Vad som är syftet att fånga avgör naturligtvis vad som är en lämplig indelning. I detta fall är indelningen av variabeln läsning relativt given, då MUCF etablerat att ungas läsning mäts utifrån andelen som läser minst varje vecka. Om intresset enbart hade varit att undersöka läsning som en fritidsaktivitet bland andra, hade ”de som läser varje månad” kunnat vara en lika lämplig indelning. Sett som en indikator för ungas levnadsvillkor kan vissa fritidsaktiviteter, som läsning och motion, kan sägas ha en bredare relevans, i termer av att regelbunden läsning kan förväntas bidra till förbättrad kunskap i språk och regelbunden motion kan bidra till ungas psykiska och fysiska hälsa. Eftersom språk är en central del i såväl grund- som gymnasieskolan blir läsningen alltså även viktig för förutsättningarna i skolan och framtida arbetsmarknadsmöjligheter (eftersom genomförd gymnasieutbildning innebär minskad risk för långvarig arbetslöshet (Engdahl & Forslund, 2015)).

4.3 Multikollinearitet i logistisk regression

Liksom för linjär regression är det i en logistisk regression viktigt att det inte finns för mycket korrelation mellan de oberoende variablerna eftersom det kan leda till att modellen inte fångar de enskilda variablernas faktiska påverkan på den beroende variabeln. Multikollinearitet kan exempelvis ta sig i uttryck av att koefficienterna i modellen förändras mycket när en variabel

läggs till. Multikollinearitet kan också ta sig i uttryck genom stor varians. Utöver dessa tecken på en problematisk korrelation mellan oberoende variabler kan också olika test användas.

4.3.1 Cramérs V

Cramérs V är ett associationsmått som är särskilt utformat för att hantera kategorivariabler där någon av variablerna kan anta fler än två värden, det vill säga då matrisen är större än 2x2. Cramérs V kan dock användas för matriser om 2x2. Måttet, V, anger hur stark associationen mellan variablerna är, och beräkans av följande ekvation:

$$V = \sqrt{\frac{\chi^2 \,/\, n}{min(I-1, J-1)}}$$

I är antalet kolumner och J är antalet rader, n är stickprovsstorleken. χ^2 är värdet på Pearsons chi^2-test (Bishop, Fienberg & Holland, 2007).

Cramérs V använder p-värdet från Pearsons chi^2-test för att avgöra om korrelation finns. Att korrelation finns innebär dock inte att den därmed alltid är problematisk. Så länge korrelationen inte är alltför stor behöver den inte utgöra ett problem.

Det finns inte konsensus vilka värden av Cramérs V som är att betrakta som höga. Lee (2016), som forskar inom medicin, menar att 0.1–0.2 är att betrakta som svag association och 0.2-0.4 som medelstor association. Dai m.fl. (2021), som också forskar inom medicin, menar att allt över 0.15 är att betrakta som en stark association. Uppgår värdet till 0.5 och högre förefaller det dock finnas konsensus om att detta är hög korrelation.

Det finns också en potentiell utmaning med bias i Cramérs V, särskilt då stickprovsstorleken är liten och matrisen är stor (det vill säga då kategorivariablerna kan anta många värden). I dessa fall tenderar Cramérs V att överskatta associationen mellan variablerna. Bergsma har utvecklat en korrigering för bias som kan användas då matrisen är större än 2x2. Korrigeringen ger störst effekt då matrisen är stor och stockprovsstorleken liten (Bergsma, 2013). Eftersom stickprovsstorleken är stor och matriserna är relativt små kommer Cramérs V användas utan korrigering i denna uppsats.

4.4 Utveckling och validering av modellen

Parametrarna för den logistiska modellen kan testas på flera sätt, såsom genom ett Wald test, ett Score test eller ett likelihoodkvot-test. Olika program använder delvis olika testmetoder för att testa signifikans, vilka sägas ge liknande resultat (Hosmer, Lemeshow & Sturdivant, 2013). STATA, vilket är programmet som används i arbetet med den här uppsatsen, använder sig av Likelihoodkvot-test för den övergripande modellen, medan Wald-test används för testning av koefficienterna.

Modellen kommer att utvecklas genom att de förklarande variablerna läggs till och utvärderas stegvis. De alternativa modellerna jämförs därefter successivt med varandra. Jämförelsen av modellerna mot varandra görs genom ett likelihoodkvot-test (se 4.4.2). I de fall modellen med ytterligare en variabel är signifikant bättre än den enklare modellen läggs ytterligare en variabel till och den nya modellen testas på samma sätt. Vid slutgiltigt val av modell kommer även AIC, BIC och pseudo R^2-värden att tas hänsyn till.

4.4.1 Wald test

Wald-testet används för att pröva koefficienterna i modellen. Värdet på koefficienten utifrån stickprovet ($\hat{\beta}_i$) prövas mot nollhypotesen att den oberoende variabeln har ingen påverkan på värdet av den beroende variabeln (β_{i0}). Wald statistikan är chi^2-fördelad och formuleras

$$W = \frac{\left(\hat{\beta}_i - \beta_{i^0}\right)^2}{V\left(\hat{\beta}_i\right)}$$

Roten av W är ett pseudo-t statistika ofta approximativt normalfördelad och då kan z-test användas för att testa om värdet på koefficienten är signifikant skilt från 0 (Davidson & MacKinnon, 1993). I Stata används ett z-test för koefficienterna.

4.4.2 Likelihoodkvot-test

Ett likelihoodkvot-test använder sig, som namnet antyder, av ration av två log likelihoodfunktioner för att jämföra modellerna. Genom att multiplicera logaritmen av det maximala värdet för log likelihoodfunktionen med -2 följer kvoten en chi^2-fördelning. Testet baseras på ration mellan de två modellernas log likelihoodfunktioner, där $l(\theta_0)$ är log likelihoodfunktionen för modellen utan förklarande variabler, vilken utgör nollhypotesen. Den jämförs mot log likelihoodfunktionen för den modell vi vill testa, $l(\theta)$).

$$\lambda_{LR} = -2\ln\left[\frac{max_{\theta\in\Theta_0} L(\theta)}{max_{\theta\in\Theta} L(\theta)}\right] = -2(l(\theta_0) - l(\theta))$$

Antalet frihetsgrader i testet avgörs av skillnaden av variabler mellan de två modellerna, vilket för test av modellen jämfört med inga variabler är samtliga variabler i modellen. För kategorivariabler med fler än två kategorier innebär detta att en frihetsgrad tillkommer för varje dummyvariabel som tillkommer.

Likelihoodkvot-test kan även användas för att jämföra modeller med olika många förklarande variabler med varandra. Då utgör den mindre modellen nollhypotesen som testas mot en modell där ytterligare förklarande variabler lagts till. Utgångspunkten är att den mindre modellen är den bästa modellen. Är p-värdet signifikant på en viss nivå förkastas nollhypotesen och den större modellen är bättre.

För att göra ett test som jämför två modeller behöver antalet observationer vara detsamma. Vid bortfall kommer antalet observationer att skilja sig mellan modellerna och det behöver hanteras för att kunna göra ett test.

4.4.3 Pseudo R^2

Pseudo R^2 är ett försök att skapa en motsvarighet till R^2, som används då variablerna som undersöks är nominala eller ordinala. Det finns flera olika sätt att ta fram ett pseudo R^2-värde. Stata använder sig av MacFaddens R^2-test. Inget av sätten motsvarar dock fullt ut R^2-värdets innebörd. Det vill säga pseudo R^2 mäter inte i vilken utsträckning förändring i den beroende variabeln förklaras av modellen. Det gör att Pseudo R^2 bör tolkas med stor försiktighet (Stata). Hosmer, Lemeshow och Sturdivant (2013) konstaterar även att R^2 generellt är låga inom logistisk regression, vilket kan gälla även om en modell är väl anpassad.

Som exempel på utmaningen att använda pseudo R^2-värdet som mått på en välanpassad modell kan ovan beskrivna studie av ungas valdeltagande nämnas. Modellerna i denna studie har flera förklarande variabler som även kommer att användas i denna uppsats och kan därför

ses som intressanta i jämförande synpunkt. Pseudo R^2-värdena för de modeller som testas i studien om valdeltagande ligger på mellan 0.05-0.11 (Abdelzadeh & Lundberg, 2020). Baserat på dessa värden är det rimligt att vänta sig att Pseudo R^2-värden för den modell som tas fram i denna uppsats också kommer att vara låg. Som ovan konstaterats tenderar R^2-värden generellt vara lägre inom logistisk regression och pseudo R^2-värden motsvarar inte heller fullt ut tenderar R^2-värden. Därför kommer värdet att rapporteras och diskuteras, men inte användas som kriterium för val av modell.

4.4.4 AIC och BIC

Utöver ett likelihood ratio-test kan Akaike Information Criterion, AIC, och Bayesian Information Criterion, BIC, användas för att jämföra modeller. AIC är ett sätt att mäta prediktionsfelet i modellen, eller hur mycket information som saknas i modellen. Ju lägre AIC, desto mindre information saknas från modellen. Som framgår nedan är AIC beroende av log likelhoodfunktionens storlek, eller modellens deviance (-2ln(L)).

$$AIC = -2\ln(L) + 2(p+1)$$

L är maxpunkten för modellens likelihoodfunktion och p är antalet parametrar (Hosmer, Lemeshow och Sturdivant, 2013). Inkluderingen av antalet parametrar minskar risken för att välja en alltför stor modell. AIC är samtidigt biased och tenderar, trots korrigeringstermen 2(p+1) att gynna stora modeller vid små urval. En korrigerad AIC, AIC_c kan användas som alternativ då stickprovet är litet eller $p/n > 10\%$ (Cryer & Chan, 2008). Eftersom stickprovet är stort bör detta inte utgöra ett problem för det aktuella underlaget.

AIC används ofta för att jämföra modeller med olika många parametrar. En modell med lägre AIC är generellt att föredra jämfört med en modell med högre AIC. Hosmer, Lemeshow och Sturdivant menar dock att en helhetsbedömning bör göras vid val av modell och att ett lägre AIC inte gör att modellen per automatik är att föredra (Hosmer, Lemeshow och Sturdivant, 2013).

Bayesian information criterion, BIC, använder liksom AIC log likelhoodfunktionens storlek, eller modellens deviance (-2ln(L)).

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

Ett lägre värde på BIC är generellt att föredra i val av modell. En skillnad mellan AIC och BIC är att BIC tenderar att bestraffa stora modeller mer än AIC. Detta kan bidra till att minska risken för att överanpassa modellen (Cryer & Chan, 2008; Stoica & Selen, 2004).

4.4.5 ROC – kurvan

ROC, reciever operating characteristic, är en graf som visar modellens förmåga att prediktera ett sant positivt utfall (y=1) mot risken för ett falskt positivt utfall. Arean under ROC-kurvan utgör ett mått på modellens prediktionsförmåga. Ett värde på 0.5 innebär att modellen prediktionsförmåga är jämförbar med att kasta ett mynt. Först vid värde på 0.7 anses prediktionsförmågan vara godtagbart bra (Hosmer, Lemeshow & Sturdivant, 2013; Gareth mfl., 2021).

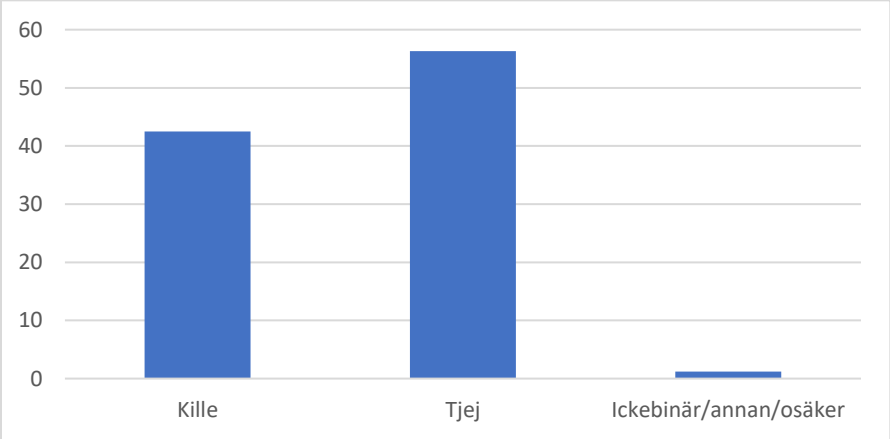
Hosmer, Lemeshow och Sturdivant (2013) rekommenderar att ROC används som mått då syftet är att särskilja två utfallsgrupper, eller skatta P(y=1). Om intresset snarare är koefficienterna eller oddskvoterna är ROC inte nödvändigtvis relevant. I kapitel 7 kommer en mer utförlig diskussion om relevansen av ROC-kurvan i uppsatsens sammanhang föras.

5 Resultat

I detta kapitel presenteras först deskriptiv statistik över datasetet, som ger en bild av sammansättningen av de svarande och hur sammansättningen ser ut i förhållande till den unga befolkningen. Därefter utvecklas och testas analysmodellen i den andra delen.

5.1 Deskriptiv statistik

Materialet består av 5186 svarande individer 16–25 år, efter att felaktiga svarande[2] och partiellt bortfall rensats. Som framgår av figur 1 är andelen killar lägre en än andelen tjejer, 42.5 respektive 56.3 procent, vilket kan jämföras med 52 killar och 48 procent tjejer i befolkningen i samma åldersgrupp 2021 (Scb, 2022). Det finns alltså en underrepresentation av killar i materialet. Därutöver har 1,2 procent svarat att en har en annan könsidentitet eller är osäkra på sin könsidentitet. Andelen med annan könsidentitet i populationen är inte känd, men utifrån uppskattningar om andelen transpersoner (vilket de flesta inom denna grupp identifierar sig som) på 0.5-2 procent av befolkningen (Scb, 2020) kan inte andelen sägas avvika avsevärt mot vad som kan förväntas.

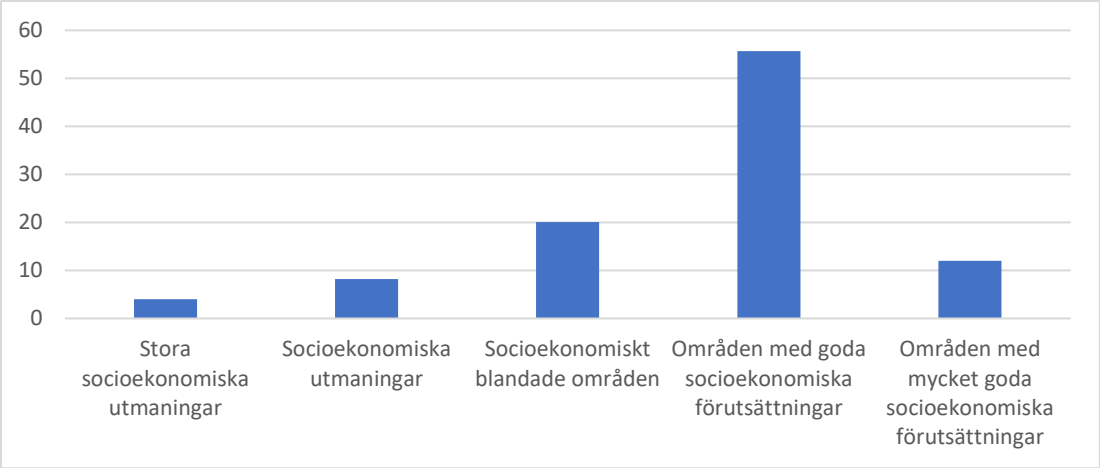


Figur 1. Andel killar, tjejer och personer med annan könsidentitet/osäkra. 16-25 år. 2021

Andelen utrikes födda bland de svarande är cirka 10 procent (se figur 3), vilket kan jämföras med 21 procent i befolkningen i åldern 16–25 år 2021 (Scb, 2022). I materialet är alltså utrikes födda underrepresenterade. Detta har också konstaterats i en bortfallsanalys som genomfördes i samband med insamlingen av materialet.

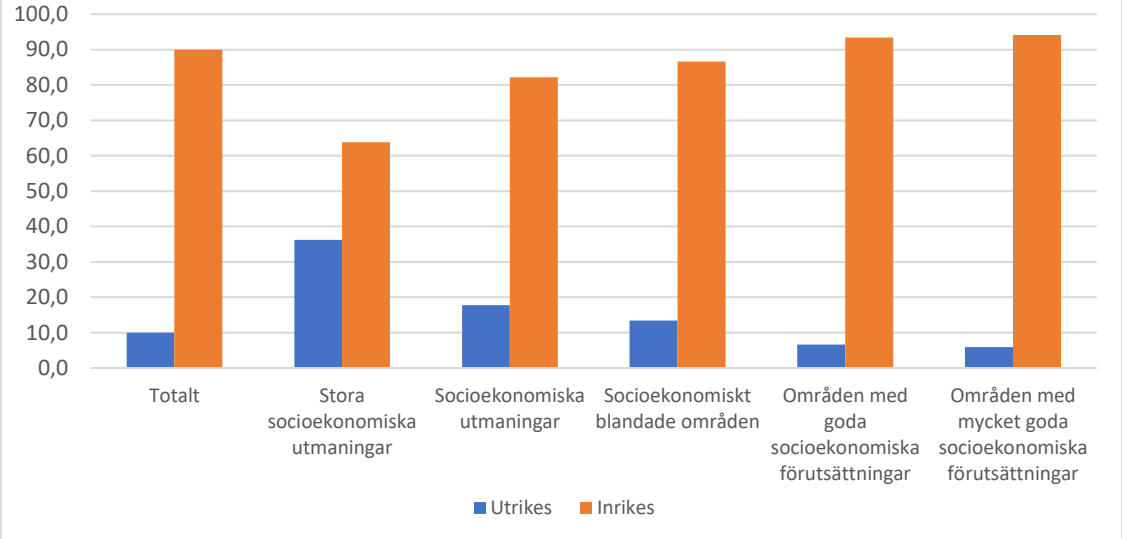
Det finns inte någon lättillgänglig information om andelen unga efter olika socioekonomiska områdestyper, men det finns information om fördelningen 2018 i Abdelzadeh och Lundbergs (2020) om ungas valdeltagande. Medan detta inte ger en exakt bild av hur fördelningen ser ut 2021, ger det en ungefärlig bild av hur fördelningen ser ut bland unga, då vi inte bör vänta oss någon stor förändring av socioekonomisk sammansättning under en så pass kort tidsperiod. 2018 bodde 5.2 procent av unga i område 1, 9.8 procent i område 2, 23.2 procent i område 3, 52.2 procent i område 4 samt 9.6 procent i område 5. Som framkommer av figur 2 följer andelarna bland de svarande samma mönster. Eftersom det inte finns information om fördelningen för 2021 går det inte att med säkerhet avgöra om det finns en underrepresentation av unga i område 1-3 eller inte.

[22] Detta rör sig om individer som enbart svarat på ett antal bakgrundsfrågor, som svarat på enkäten under orimligt kort tid eller individer där svarsmönstren avviker väsentligt från vad som kan förväntas. Rensningen är genomförd av MUCF, sedan tidigare. Det finns utöver detta ett partiellt bortfall



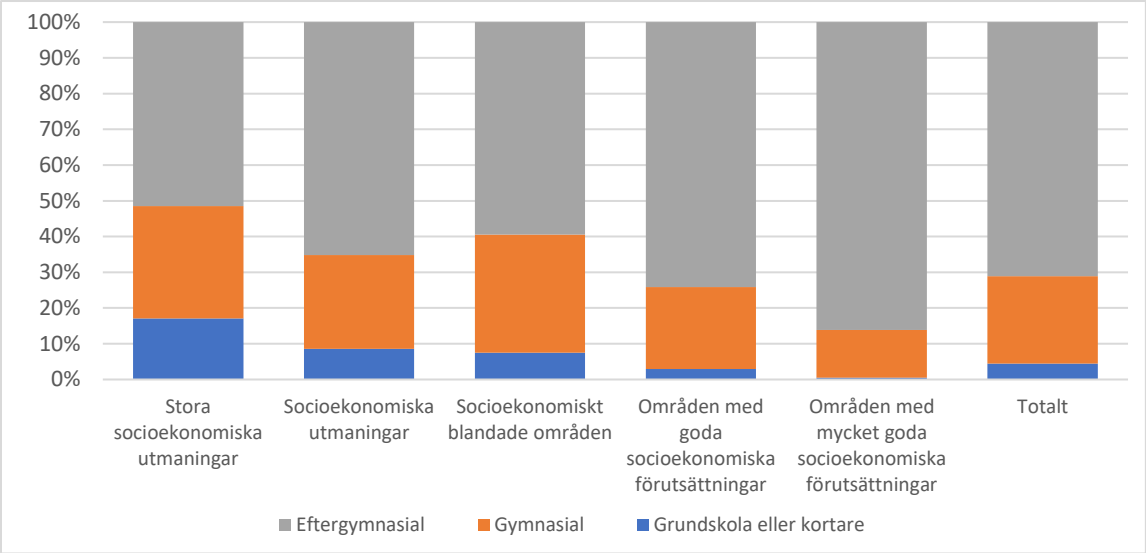
Figur 2. Andel unga efter områdestyp.16-25 år. 2021.

Som framkommer av figur 3 är andelen unga utrikesfödda lägre för varje område med bättre socioekonomiska förutsättningar, vilket är väntat, eftersom andelen utrikesfödda är högre i områden med sämre socioekonomiska förutsättningar (Boverket, 2023).



Figur 3. Andelen unga inrikes respektive utrikes födda, efter områdestyp. 16-25 år. 2021.

En bakgrundsvariabel som är relaterad till socioekonomisk områdestyp är föräldrarnas utbildningsnivå. Som framkommer av figur 3 har de flesta som svarat på enkäten minst en förälder med eftergymnasial utbildning, medan bara en liten andel har föräldrar med högst grundskoleutbildning.



Figur 4. Föräldrars högsta avslutade godkända utbildning, efter områdestyp. 16-25 år. 2021.

Källa: Nationella ungdomsenkäten. MUCF.

5.1.1 Läsning bland olika grupper av unga

Andelen unga som läser böcker varje vecka 2021 är 29 procent. Som framkommer av tabellerna nedan är läsningen inte jämnt fördelad bland unga. Exempelvis är det en högre andel unga i områden med socioekonomiska utmaningar (2) som läser jämfört med unga i områden med blandade (3), goda (4) eller mycket goda (5) socioekonomiska förutsättningar. Det är också klart vanligare att tjejer och unga med annan könsidentitet läser jämfört med killar och att utrikes födda i högre grad läser varje vecka jämfört med inrikes födda unga. Detta har också framkommit i en tidigare rapport av MUCF (2023).

Tabell 4. Andel unga som läser varje vecka, efter områdestyp.

	Läser varje vecka	**Konfidensintervall (95%)**	
Område med stora socioekonomiska utmaningar	34%	27.56%	40.34%
Område med socioekonomiska utmaningar	38%	33.08%	42.26%
Socioekonomiskt blandade områden	28%	24.94%	30.37%
Område med goda socioekonomiska förutsättningar	29%	27.71%	31.03%
Område med mycket goda socioekonomiska förutsättningar	28%	24.78%	31.85%

Tabell 5. Andel unga som läser varje vecka, efter födelseland.

	Läser varje vecka	Konfidensintervall (95%)	
Inrikes födda	29%	27.67%	30.27%
Utrikes födda	36%	32.33%	40.60%

Tabell 6. Andel unga som läser varje vecka, efter kön.

	Läser varje vecka	Konfidensintervall (95%)	
Kille	25%	23.26%	26.87%
Tjej	33%	31.12%	34.5%
Annan/ickebinär/osäker	49%	37.14%	61.35%

Som framkommer av tabellen nedan är det också så att unga med minst en förälder med eftergymnasial utbildning läser i högre grad än unga med föräldrar med kortare utbildning.

Tabell 7. Andel unga som läser varje vecka, efter föräldrars utbildningsnivå.

	Läser varje vecka	Konfidensintervall (95%)	
Grundskola/gymnasial/ingen	25%	23.08%	27.47%
Eftergymnasial	32%	30.04%	33.04%

I uppsatsen finns även en ambition om att undersöka samvariationen mellan ungas biblioteksbesök och ungas läsning. Som framkommer av tabell 8 finns en tydlig samvariation mellan att besöka bibliotek och att läsa varje vecka. Bland de som sällan eller aldrig besöker bibliotek är det en klar minoritet som läser varje vecka.

Tabell 8. Andel som läser varje vecka efter biblioteksbesök.

	Läser varje vecka	Konfidensintervall	
Besöker bibliotek varje dag	63%	43.77%	78.78%
Besöker bibliotek varje vecka	64%	58.11%	69.97%
Besöker bibliotek varje månad	54%	50.49%	56.88%
Besöker bibliotek varje år	29%	27.03%	31.13%
Besöker aldrig bibliotek	15%	13.63%	16.69%

6 Analys

Modellen för att förklara ungas läsning utvecklas genom att förklarande variabler läggs till efter hand och prövas genom likelihoodkvot-test (LR-test) mot den föregående modellen med färre förklarande variabler. I det fall modellen genom LR-test är signifikant förkastas hypotesen om att den mindre modellen är det bästa preliminärt, och ytterligare en förklarande variabel läggs till och testas på samma vis. Därutöver kommer multikollinearitet mellan de oberoende variablerna undersökas. För slutgiltig bedömning av modellen kommer hänsyn även tas till AIC och BIC, där ett lägre värde indikerar på en bättre modell. ROC används ofta för att bedöma en modell, där ett högre värde på ROC indikerar på en modell med bättre prediktion. ROC kommer att undersökas. En diskussion om måttets relevans förs i kapitel 7.

6.1.1 Analys av bortfall

Samtliga modeller utgår från ett rensat dataset där observationer som innehåller bortfall uteslutits. Samtliga modeller har dock också testats med observationerna där detta har varit möjligt. Det vill säga, då de saknade svaren inte påverkar den aktuella modellen.

För de flesta variabler består bortfallet enbart av ett mindre antal individer. För variabeln föräldrars utbildningsnivå medför det ett bortfall på totalt 359 individer, varför det är särskilt viktigt att undersöka närmare.

Osbourne (2015) betonar vikten av att undersöka observationerna med bortfall noggrant samt understryker att radering av observationer inte bör göras utan eftertanke. Att observationerna som saknas inte kan anses vara slumpmässigt saknade innebär att resultatet i analysen riskerar bias. I fallet med föräldrarnas utbildning innebär bland annat radering av data att modellen inte kan säga något om läsning bland unga som av något skäl saknar kunskap om sina föräldrars utbildning.

Den största delen av bortfallet, 254 individer, består i unga som svarat att de inte vet vad föräldrarna har för utbildning. För att undersöka dessa individer närmare görs en separat logistisk regressionsmodell där dessa inkluderas som en kategori. Analysen visar att denna grupp inte signifikant skiljer sig från unga som känner till föräldrars utbildningsnivå, oavsett föräldrarnas utbildningsnivå. Sett till koefficientera liknar de framför allt unga med föräldrar som har kortare än eftergymnasial utbildning (se bilaga 1).

För att undersöka hur partiella bortfall har modellerna tagits fram både med och utan observationer med partiellt. Resultatet av analyserna redovisas i bilaga 1. Som väntat är Likelihood-funktionens värden något bättre och konfidensintervallen för koefficienterna mindre. Inga tester ger dock några andra resultat, och z-värdena förändras bara lite, då observationerna läggs till. Detta tyder på att bortfallet hos en viss variabel inte avsevärt påverkar resultatet för de andra oberoende variablerna.

6.2 Modell 1

Modell 1 inkluderar enbart kön. I modellen jämförs tjejer och personer med annan könsidentitet med killar. Som framkommer av LR-testet och dess p-värde är modellen signifikant vid jämförelse mot en modell med enbart intercept. Pseudo R^2 är mycket lågt. Att värdet är lågt är dock inte ovanligt vid tester som undersöker samhällsvetenskapliga företeelser.

Tabell 9. Modell 1. Läser varje vecka efter kön.

Log likelihood	LR-test	p-värde	Pseudo R^2	Antal obs.
-3130.6224	47.64	0.000*	0.0076	5186

6.3 Modell 2

Modell 2 inkluderar även födelseland, där inrikes födda unga jämförs med utrikes födda unga. Modellen som helhet samt koefficienterna är signifikanta med p-värden på 0.000.

Tabell 10. Modell 2. Läser varje vecka efter kön och födelseland

Log likelihood	LR-test	p-värde	Pseudo R^2	Antal obs.
-3124.4051	60.07	0.000*	0.0095	5186

Likelihoodkvot-testet som jämför modell 1 och 2 är signifikant med ett p-värde på 0.0005, vilket ger en stark indikation på att modellen med både kön och födelseland är en bättre modell än modell 1 (se tabell 14). AIC och BIC är något lägre för modell 2 jämfört modell 1, men skillnaden är liten (se tabell 15). Baserat på detta används modell 2 som utgångspunkt för vidare utveckling.

6.4 Modell 3 och 4

Modell 3 och 4 inkluderar båda områdestyp som tredje variabel, med använder olika kategorier som jämförelse för de dummyvariablerna som tillsammans utgör områdestyperna. Eftersom modell 3 och 4 innehåller samma variabler ger likelihoodkvot-test samma resultat för båda modellerna. Som framkommer i tabell 11 är modellen som helhet signifikant jämfört med enbart interceptet. Värdet på LR-testet Pseudo R^2-värdet är något högre jämfört med modell 1 och 2.

Ingen av områdestyperna 2-5 signifikant skilda från områdestyp 1, områden med stora socioekonomiska skillnader. Däremot visar modell 4, som jämför områdestyperna med områdestyp 5, områden med mycket goda socioekonomiska förutsättningar, att områdestyp 2, områden med socioekonomiska utmaningar, är signifikant skilt från områdestyp 5. Eftersom modell 4, till skillnad från modell 3, tydliggör på vilket sätt områdestyperna påverkar läsning kommer områdestyp 5 vara jämförelsekategori fortsatt utveckling av modellen. Alla Likelihoodkvot-tester kommer dock ge samma resultat för modell 3 och 4, eftersom de innehåller samma information i grunden.

Tabell 11. Modell 3 och 4. Läser varje vecka efter kön, födelseland och områdestyp.

Log likelihood	LR-test	p-värde	Pseudo R^2	Antal obs.
-3117.0219	74.84	0.000*	0.0119	5186

Likelihoodkvot-testet jämför modell 4 med modell 2 ger ett värde på 14.88 och ett p-värde på 0.0052, vilket är signifikant (se tabell 14). Nollhypotesen att den bästa modellen enbart inkluderar födelseland och kön förkastas alltså. AIC-värdet för modell 4 är något lägre än för modell 2, men skillnaden är liten, 6250, jämfört med 6256. Däremot är BIC-värdet högre för modell 4 jämfört med modell 2 (se tabell 15). AIC- och BIC-värdena ger alltså olika indikation på vilken modell som är bäst.

6.5 Modell 5

I den 5 modellen läggs föräldrars utbildningsnivå till modellen. Variabeln är en dummyvariabel där unga med minst en förälder med eftergymnasial utbildning jämförs med unga med föräldrar med kortare utbildning.

LR-testet för modell 5 som helhet jämfört med en modell med enbart intercept är signifikant. För de variabler som inkluderats i tidigare modeller har mönster och signifikans inte ändrats. Koefficienten för föräldrars utbildningsnivå är signifikant med ett p-värde på 0.000.

Tabell 12. Modell 5. Läser varje vecka efter kön, födelseland, områdestyp och föräldrars utbildningsnivå.

Log likelihood	LR-test	p-värde	Pseudo R^2	Antal obs.
-3106.4278	96.02	0.000*	0.0152	5186

6.6 Modell 6

LR-testet för modell 6 som helhet jämfört med en modell med enbart intercept är signifikant. För de variabler som inkluderats i tidigare modeller har mönster och signifikans inte ändrats. Dock är p-värdet för utrikes födda på 0.028 något högre jämfört med tidigare modeller då värdet var mellan 0.000 och 0.002. Biblioteksbesöksvariabeln är signifikant med ett p-värde på 0.000.

Tabell 13. Modell 6. Läser varje vecka efter kön, födelseland, områdestyp, föräldrars utbildningsnivå och biblioteksbesök.

Log likelihood	LR-test	p-värde	Pseudo R^2	Antal obs.
-3039.8556	229.51	0.000*	0.0364	5186

Variabel	Oddskvot	Std. Err.	p-värde
Kön: Tjej	1.446595	0.064	0.000*
Kön: Annan	2.679147	0.262	0.000*
Födelseland: Utrikes	1.255006	0.103	0.028*
Områdestyp: 1	1.282652	0.182	0.171
Områdestyp: 2	1.496525	0.139	0.004*
Områdestyp: 3	1.023884	0.117	0.840
Områdestyp: 4	1.096474	0.100	0.359
Föräldrars utb.nivå: Eftergymnasial	1.397264	0.072	0.000*
Besöker bibliotek varje vecka	4.392125	0.131	0.000*

LR-testet som jämfört modell 5 och modell 6 är signifikant med test-värdet 133.48 och p-värdet 0.000 (se tabell 14). Nollhypotesen att modell 5 är den bättre modellen förkastas alltså. AIC för modell 6 är 6230.856 jämfört med 6240.063 för modell 5, alltså något lägre. Även BIC för modell 6 är lägre jämfört med modell 5, 6164.908. jämfört med 6289.839 (se tabell 15). Detta tillsammans med att såväl pseudo R^2 som LR-testet för hela modellen är högre. Även om pseudo R^2-värdet 0.036 är lågt är det högre jämfört med 0.015. Av dessa modeller förefaller alltså modell 6 vara att föredra.

Tabell 14. LR-test, jämförelse av modellerna.

	M1 och M2	M2 och M4	M4 och M5	M5 och M6
LR-test	12.43	14.77	21.19	133.48
P-värde	0.0004	0.0052	0.0000	0.0000

Tabell 15. AIC och BIC för samtliga modeller.

	Modell 1	Modell 2	Modell 3/4	Modell 5	Modell 6
AIC	6267.245	6256.81	6250.044	6230.856	6099.371
BIC	6286.906	6283.025	6302.473	6289.839	6164.908

Som beskrivs i metodkapitlet är värdet på koefficienterna inte helt lätta att tolka rakt av för att förstå den oberoende variabelns påverkan på den beroende variabeln. Istället kan oddskvoter kan användas för att tydliggöra variabelns påverkan (se tabell 13). Eftersom alla variabler är kategorivariabler och hanteras som dummyvariabler i modellen ska oddskvoten tolkas som förändring i sannolikhet att en ung person läser varje vecka om personen har en viss egenskap jämfört med den andra egenskapen. Som framkommer av tabell 13 har de flesta variablerna som är signifikanta relativt modest storlek på oddskvoterna. Exempelvis finns en positiv samvariation mellan att vara tjej och läsa varje vecka jämfört med en kille. Två variablers oddskvoter sticker ut, unga med annan könsidentitet/osäkra och unga som besöker bibliotek. Oddskvoten för unga med annan könsidentitet är 2.7 jämfört med killar[3] och oddskvoten för unga som besöker bibliotek varje vecka är 4.4 jämfört med unga som besöker bibliotek mer sällan.

6.7 Multikollinearitet

För att kunna bedöma hur väl modell 6 fungerar undersöks eventuell multikollinearitet.

Tabell 16. Cramérs V

	Födelseland	Områdestyp	Kön	Föräldrars utb.nivå	Biblioteksbesök
Födelseland	1				
Områdestyp	0.2315*	1			
Kön	0.0131	0.0237	1		
Föräldrars utb.nivå	-0.0224	0.1919*	0.0481*	1	
Biblioteksbesök	0.0852*	0.0744*	0.0383*	-0.0061	1

Kommentar. Tal markerade med * innebär att Pearson chi2-test är signifikant skilt från 0.

Cramérs V visar på tal under 0.05 för association mellan kön samtliga andra variabler, mellan födelseland och föräldrars utbildningsnivå samt mellan biblioteksbesök och föräldrars utbildningsnivå. Cramérs V är under 0.1 för biblioteksbesök och områdestyp samt biblioteksbesök och födelseland. Pearsons chi2-test visar att association finns för ett par av dessa, är associationen att betrakta som låg. För de variabler där associationen var signifikant skild från 0 gjordes en närmare undersökning av hur associationen såg ut. Associationen

[3] Antalet individer i denna kategori är dock få.

mellan föräldrars utbildningsnivå och kön är intressant, eftersom det i princip enbart är en skev fördelning för unga med annan könsidentitet/osäkra som bidrar till denna association.

Mellan områdestyperna och födelseland samt föräldrars utbildningsnivå är värdena för Cramérs V högre. Eftersom koncentrationen av en viss utbildningsnivå i den vuxna befolkningen är något som utgör en del i indexet som definierar områdestyperna är detta inte oväntat. Inrikes och utrikes födda är även ojämnt fördelade mellan de olika områdestyperna. Det är alltså inte heller förvånande att se att värdet är högre här. Som konstaterats i metodkapitlet finns inte konsensus om värden på 0.19 respektive 0.23 är att betrakta som högt eller inte. Värdena i sig kan alltså tolkas både som problematiska och oproblematiska. För att bedöma huruvida det finns en problematisk korrelation mellan dessa variabler kan en jämförelse av koefficienterna i modell 4, 5 och 6 vara till stöd för bedömning. En stor förändring av koefficienterna eller ett en stor varians/medelfel mellan koefficienterna i modellerna är en indikation på multikollinearitet.

Tabell 17. Koefficienternas värde för de olika modellerna

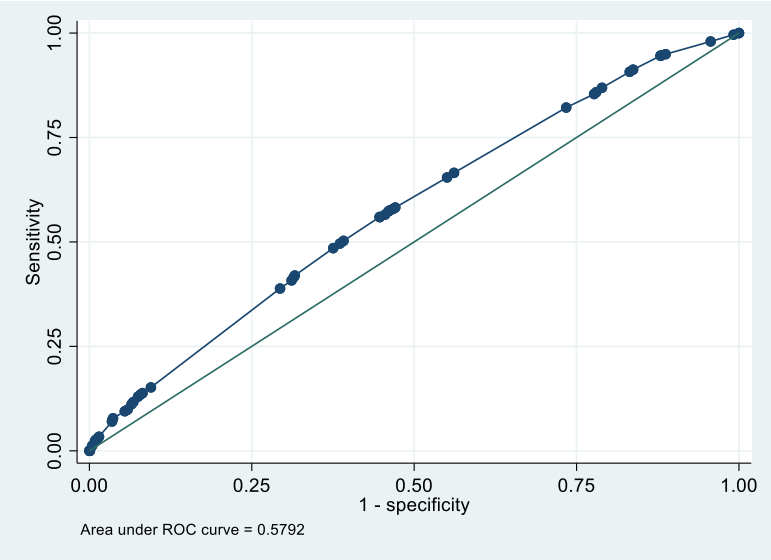
		Modell				
		1	**2**	**4**	**5**	**6**
Kön	**Tjej**	0.3805	0.3824	0.3851	0.3839	0.3692
	Ickebinär/osäker/annan	1.0657	1.0817	1.0968	1.0407	0.9854
Utrikes	**Utrikes**		0.3483	0.3149	0.3109	0.2271
Områdestyp	**1**			0.1702	0.2817	0.2489
	2			0.4088	0.4756	0.4031
	3			-0.0464	0.0348	0.0236
	4			0.0633	0.1009	0.0921
Föräldrars utb.	**Hög utbildning**				0.3231	0.3345
Biblioteksbesök	**Varje vecka**					1.4798

Det sker en liten förändring av koefficienten för personer med annan könsidentitet mellan modell 4 och 5, samt mellan modell 5 och 6. Förändringen är dock liten. Medelfelet är i princip konstant på cirka 0.26 i samtliga modeller.

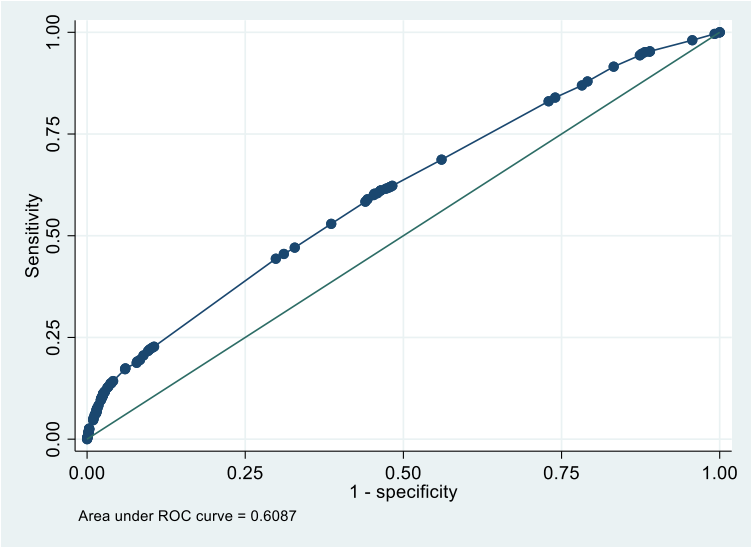
Det ser viss förändring av koefficienten för utrikes födda mellan modell 5 och 6, från 0.31 till 0.23. Medelfelet för koefficienten är konstant på 0.10 mellan modellerna. Mellan modell 4 och 5 sker en viss förändring av koefficienten för områdestyp 1, områdestyp 2 och områdestyp 3. Medelfelen är i princip oförändrade mellan modellerna. Mellan modell 5 och 6 sker en mindre förändring av koefficienten för områdestyp 2. Samtliga förändringar av koefficienterna är mindre än medelfelet för respektive koefficient. Ingen av dessa förändringar påverkar heller slutsatserna som dras om respektive variabel. Den sammantagna bedömningen är därför att multikollineariet inte utgör ett problem i modellen.

6.8 Prediktionsförmåga

Nedan figurer visar ROC-kurvor och värden för modell 5 och 6. Syftet att även undersöka modell 5 är för att kunna jämfört modell 6 mot en enklare modell.



Figur 5. ROC-test för modell 5.



Figur 6. ROC-test för modell 6.

Som figur 4 och 5 visar är arean under ROC-kurvan något större för modell 6 jämfört med modell 5, 0.6087 jämfört med 0.5792. Med andra ord har ingen av modellerna en bra prediktionsförmåga. Det vill säga, utifrån modellen kommer det inte vara möjligt att med någon större säkerhet förutsäga, baserat enbart på dessa variabler, om en individ med en viss uppsättning egenskaper läser varje vecka.

7 Avslutande diskussion

Syftet med denna uppsats har varit att undersöka hur födelseland och socioekonomisk områdestyp påverkar andelen unga som läser varje vecka och vilken metod som är bäst lämpad för att undersöka detta. Syftet har även varit att undersöka om en fördjupad analys bidrar till en bättre förståelse för skillnader i läsning och vilka slutsatser som är möjliga att dra utifrån en sådan analys.

Den modell som bäst förklarar ungas läsning har fem förklarande variabler; kön, födelseland, områdestyp, föräldrars utbildningsnivå och biblioteksbesök. Allra störst positiv samvariation med läsning förefaller vara att besöka bibliotek varje vecka, följt av att vara ickebinär/osäker på sin könsidentitet. Även att ha minst en högskoleutbildad förälder, att vara tjej och att vara utrikes född påverkar läsningen positivt. Resultatet för områdestyperna är något oväntat, att det är troligare att unga i områden med socioekonomiska utmaningar har större sannolikhet att läsa varje vecka jämfört med unga i områden med mycket goda socioekonomiska förutsättningar. Samtidigt bidrar resultatet till att klargöra en av de frågor som föranledde val av ämne för denna uppsats. Är det så att överrepresentationen av utrikes födda i områden med socioekonomiska utmaningar osynliggör att unga i dessa områden läser i lägre utsträckning. Resultatet pekar på att så inte är fallet. Det finns däremot andra faktorer som kan påverka resultatet kopplat till socioekonomiskt bostadsområde, vilket kommer att diskuteras nedan.

7.1 Hänsyn till syftet viktigt för att välja rätt analysmodell

Eftersom läsning i det aktuella datamaterialet är en kategorivariabel på en ordinalskala hade tre olika typer av logistisk regressionsanalys varit möjlig att använda; binär logistisk regression, multinominal logistisk regression samt ordinal logistisk regression. Eftersom syftet med analysen framför allt är att undersöka områdestypernas påverkan har det varit viktigt att, i det fall det är möjligt, genomföra analysen med samtliga fem områdestyper. Detta gör att ordinal eller multinominal logistisk regression är det bäst lämpade valet, enbart om antalet observationer är tillräckligt stort inom samtliga kategorier av läsfrekvens för alla områdestyper. Trots att antalet observationer i materialet är stort, 5186, är antalet inom vissa kategorier betydligt mindre. Antalet observationer då samtliga kategorier av läsfrekvens används blir för områdestyp 1 lägre än 10 för vissa kategorier av läsfrekvens, vilket medfört en stor osäkerhet i skattningarna och påverkat vilka slutsatser som är möjliga att dra. Detta gör att en binär logistisk modell är den bäst lämpade metoden för det aktuella syftet.

7.2 Troligt bortfall av unga med föräldrar med kortare utbildning minskar resultatets tillförlitlighet

Resultatet av analysen ligger i linje med resultat från rapporten *Ung idag 2023 En fördjupad bild av ungas fritid* (MUCF, 2023).

Vid närmare undersökning av fördelningen av föräldrars utbildningsnivå mellan de olika områdestyperna finns dock anledning att fundera över tillförlitligheten i resultaten. Oavsett områdestyp utgör unga med minst en högskoleutbildad förälder majoritet. Detta trots att områden med socioekonomiska utmaningar präglas av an hög andel med lägre utbildningsnivå. Det vore alltså rimligt att förvänta sig att en låg andel av föräldrarna har en högre utbildning. Medan det finns en skillnad mellan områdestyperna är det alltså tveksamt om dessa unga kan anses representativa för unga i områden med socioekonomiska utmaningar, när det gäller föräldrars utbildningsnivå. Eftersom modellen tydliggör att utbildningsnivå är betydelsefullt för läsning bör alltså resultatet kopplat till socioekonomi

tolkas med försiktighet. Utifrån dessa resultat kan den logistiska regressionsanalysen sägas bidra till en bättre förståelse för skillnader, eller avsaknad av skillnader, i läsning mellan olika grupper av unga. Analysen bidrar också till en ökad förståelse för konsekvenser av bortfall som inte enkelt kan mätas, då vi inte känner till det exakta antalet i populationen. Exempelvis bör resultat från enkätundersökningen som visar på att det inte finns skillnader mellan områdestyper tolkas med försiktighet generellt, särskilt i de fall tidigare studier visat att socioekonomi är betydelsefullt eller det finns skillnader kopplat till föräldrars utbildningsnivå.

7.3 Enkätundersökningar medför risk för olika tolkningar som leder till mätfel

En enkätundersökning medför alltid risk att de som svarar på frågorna svarar ”fel” i förhållande till vad den som genomför eller analyserar enkäten har tänkt sig. Detta gäller fel som sker av misstag, det vill säga slumpmässiga fel. Ett annat typ av ”felsvar” att ta hänsyn till i sammanhanget är risken för att de svarande tolkar svarsalternativen för hur ofta de läser på olika sätt, vilket under vissa förutsättningar skulle kunna snedvrida resultaten. Om en person exempelvis läser varannan vecka, är det möjligt att personen tycker svarsalternativet ”varje vecka” bäst beskriver hur ofta hen läser. Det är dock också möjligt att personen resonerar att svarsalternativet ”varje månad” bäst beskriver hur ofta hen läser. Beroende på hur personen bedömer svarsalternativen kan individen falla inom de två binära alternativen för den beroende variabeln. Att denna typ av olika tolkningar förekommer får bedömas som troligt. Om det inte finns skillnader mellan grupper och de olika tolkningarna fördelar sig lika mellan olika grupper bör detta inte ha någon större påverkan på resultatet. Men då skillnader finns är det tänkbart att vissa grupper i högre utsträckning kan ställas inför just den tolkningssituation som gör att de hamnar i den ena eller andra svarsgruppen.

7.4 ROC-kurvan och Pseudo R^2 som mått på en välanpassad modell

Ofta undersöks ROC-kurvan för att bedöma en modell i logistisk regression. Syftet med ROC-kurvan är att avgöra modellens prediktionsförmåga. R^2-värdet, eller i detta fall pseudo R^2-värdet, används också generellt inom regression för att visa på i vilken utsträckning modellen förklarar utfallet på den beroende variabeln. Som Hosmer, Lemeshow och Sturdivant diskuterar krävs viss reflektion kring hur båda dessa mått ska användas och tolkas inom logistisk regression.

Som modellen visar samvarierar läsning med kön, socioekonomi, föräldrars utbildningsnivå, ungas födelseland och biblioteksbesök. Pseudo R^2-värdet för modellen är dock lågt, vilket också var väntat. I sammanhanget bör det nämnas att pseudo R^2-värdet, som tidigare nämnts, måste tolkas med stor försiktighet. Generellt kan också sägas att ett lågt R^2-värde är inte ovanligt inom logistisk regression. Att R^2-värdet generellt är lägre inom logistisk regression jämfört med linjär regression förefaller intuitivt rimligt. En företeelse är ofta mer komplex än vad som kan beskrivas av två binära kategorier. Eftersom den beroende variabeln enbart kan anta två värden, blir den ofta ett trubbigt mått på det som ska undersökas. Exempelvis finns det ett stort utrymme mellan att läsa varje vecka och att aldrig läsa på fritiden. För den binära beroende variabeln faller dock hela det spektrumet inom samma kategori.

Resultatet av grafen över ROC-kurvan visar även att modellen inte har en god prediktionsförmåga. Är detta då ett problem för modellen? Syftet med regressionsanalysen är att undersöka om skillnader finns mellan grupper av unga på samhällsnivå. Syftet är alltså inte att förutsäga om en ung person läser baserat på en uppsättning egenskaper, utan att bedöma om skillnader mellan grupper av unga finns. Om skillnader finns indikerar de på att det finns

skillnader i levnadsvillkor mellan grupper av unga som kräver insatser. Att en viss grupp unga har 50 procent högre sannolikhet att läsa, jämfört med en annan grupp är ett intressant och viktigt resultat i sammanhanget. Därför blir prediktionsförmågan i modellen av mindre betydelse. Arean under ROC-kurvan bör i det här fallet alltså inte användas för att utesluta en modell som bidrar till att belysa skillnader i läsning mellan olika grupper av unga. Däremot understryker resultatet vad även Pseudo R^2-värdet indikerar. Nämligen att en ung persons preferenser och intressen är betydligt mer komplexa processer än vad som kan förklaras av denna typ av modell.

Referenser

Abdelzadeh, A. & Lundberg, E. (2020). *Ungas röst. En studie om ungas valdeltagande 2018 och deras egna tankar om att delta i val.* Växjö: Myndigheten för ungdoms- och civilsamhällesfrågor.

Barone, C., Fougere, D. & Pin, C. (2021). Social Origins, Shared Book Reading, and Language Skills in Early Childhood: Evidence from an Information Experiment. *European Sociological Review*, 37:1, 18–31.

Bergsma, W. (2013). "A bias-correction for Cramér's V and Tschuprow's T". *Journal of the Korean Statistical Society*. 42:3. 323-328.

Bishop, Y., Fienberg, S. E. & Holland, P. W. (2013). *Discrete Multivariate Analysis. Theory and Application*. Springer.

Boverket (2022). *Så mäter och följer du segregation. Användarhandbok för Segregationsbarometern*. Karlskrona: Boverket.

Boverket (2023). *Boendesegregationens utveckling och mekanismer. Årsrapport 2023 om den socioekonomiska bostadssegregationens utveckling i Sverige.* Rapport 2023:23. Karlskrona: Boverket.

Cryer, J. D. & Chan, K.-S. (2008). *Time Series Analysis with Applications in R*. 2 ed., Springer.

Dai, J., Teng, L., Zhao, L., & Zou, H. (2021). "The combined analgesic effect of pregabalin and morphine in the treatment of pancreatic cancer pain, a retrospective study". *Cancer Medicine*. 10:5. 1738–1744.

Davidson, Russell; MacKinnon, James G. (1993). "The Method of Maximum Likelihood: Fundamental Concepts and Notation". *Estimation and Inference in Econometrics*. New York: Oxford University Press

Ee Loh C., Sun B. & Majid S. (2020). "Do girls read differently from boys? Adolescents and their gendered reading habits and preferences". *English in Education*. 54:2. 174-190.

Engdahl, M. & Forslund, A. (2015). *En förlorad generation? – en ESO-rapport om ungas etablering på arbetsmarknaden*. Stockholm: Fritzes.

Hosmer, D.W., Lemeshow, S. & Sturdivant, R.X. (2013). *Applies logistic regression. 3 ed.*

Kurnaz, A. & Pursun, T. (2022). An Analysis of the Reading Motivation of Secondary School Students in Relation to Various Variables. *Research in Pedagogy*. 12:1. 29-44.

Lathouras, M., Westerveld, M. F. & Trembath, D. (2019). "Longitudinal reading outcomes in response to a book-based, whole class intervention for students from diverse cultural, linguistic and socio-economic backgrounds", *Australian Journal of Learning Difficulties*. 24:2. 147-161.

Lee, D. K. (2016). "Alternatives to P value: Confidence interval and effect size". *Korean Journal of Anesthesiology*, 69:6. 555–562.

Lepper, C., Stang-Rabrig , J. & McElvany, N. (2022). ”Gender differences in reading: Examining text-based interest in relation to text characteristics and reading comprehension”. *Learning and Instruction.* Vol. 82.

McGinnity, F. m.fl. (2022). “Understanding differences in children’s reading ability by social origin and gender: The role of parental reading and pre- and primary school exposure in Ireland”. *Research Social Stratification and Mobility.* Vol. 81.

MUCF (2023). *Ung idag 2023:2. En fördjupad bild av ungas fritid.* Växjö: Myndigheten för ungdoms- och civilsamhällesfrågor.

MUCF (2021). ”Läser böcker”. Kultur och fritid. Ungidag.se. https://www.ungidag.se/indikator/kultur_och_fritid/laeser-boecker. Hämtad 2023-12-30.

Osbourne, J. 2015. *Best practices in logistic regression.* Los Angels: SAGE Publications

Regeringens proposition 2013/14:191. *Med fokus på unga – en politik för goda levnadsvillkor, makt och inflytande*.

Regeringens skrivelse 2020/21:105. *Ungdomspolitisk skrivelse*.

Scb (2022). Statistikdatabasen. Folkmängd efter ålder, kön, födelseregion och år. https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__BE__BE0101__BE0101E/InrUtrFoddaRegAlKon/table/tableViewLayout1/ (Hämtad 15 november 2023)

Scb (2020). Stödmaterial för att inkludera transpersoner i enkäter och undersökningar. https://www.scb.se/globalassets/stodmaterial-for-att-inkludera-transpersoner-i-enkater-och-undersokningar.pdf. Hämtad 2023-12-17.

Smith, E. & Reimer, D. (2023). “Understanding gender inequality in children's reading behavior: New insights from digital behavioral data”. *Child development.* *https://srcd.onlinelibrary.wiley.com/doi/10.1111/cdev.14001*. Hämtad 2023-12-30.

Stata. “Logistic Regression Analysis. Stata Annotated Output.” https://stats.oarc.ucla.edu/stata/output/logistic-regression-analysis/. Hämtad 2023-12-31.

Statens medieråd (2023). *Ungar och medier 2023. En statistisk undersökning av ungas medievanor och attityder till medieanvändning.* Stockholm: Statens medieråd.

Stoica, P. & Selen, Y. (2004). Model-order selection: a review on information criterion rules. IEEE Signal Processing Magazine. 21:4. 36-47.

Bilaga 1. Utskrifter från STATA

Logistisk regression

Modell 1.

Tabell 18. Modell 1, inklusive observationer med partiellt bortfall

```
Iteration 0:   log likelihood = -3361.8514
Iteration 1:   log likelihood = -3336.1276
Iteration 2:   log likelihood = -3335.9716
Iteration 3:   log likelihood = -3335.9716

Logistic regression                                     Number of obs =  5,555
                                                        LR chi2(2)    =  51.76
                                                        Prob > chi2   = 0.0000
Log likelihood = -3335.9716                             Pseudo R2     = 0.0077
```

F12G_läser_böcker_v	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
kon_egenid						
Tjej	.3752184	.0608957	6.16	0.000	.255865	.4945718
Icke-binär/Osäker/~n	1.109147	.2436528	4.55	0.000	.6315968	1.586698
_cons	-1.109147	.0471573	-23.52	0.000	-1.201574	-1.016721

Tabell 19. Modell 1, exklusive observationer med partiellt bortfall

```
Iteration 0:   log likelihood = -3154.4401
Iteration 1:   log likelihood = -3130.7424
Iteration 2:   log likelihood = -3130.6224
Iteration 3:   log likelihood = -3130.6224

Logistic regression                                     Number of obs =  5,186
                                                        LR chi2(2)    =  47.64
                                                        Prob > chi2   = 0.0000
Log likelihood = -3130.6224                             Pseudo R2     = 0.0076
```

F12G_läser_böcker_v	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
kon_egenid						
Tjej	.3805075	.0630185	6.04	0.000	.2569936	.5040214
Icke-binär/Osäker/~n	1.065655	.2567572	4.15	0.000	.5624203	1.56889
_cons	-1.097404	.0491547	-22.33	0.000	-1.193745	-1.001062

Modell 2

Tabell 20- Modell 2, inklusive observationer med partiellt bortfall

```
Iteration 0:   log likelihood = -3342.6429
Iteration 1:   log likelihood = -3308.2462
Iteration 2:   log likelihood = -3308.0473
Iteration 3:   log likelihood = -3308.0473

Logistic regression                                     Number of obs =  5,525
                                                        LR chi2(3)    =  69.19
                                                        Prob > chi2   = 0.0000
Log likelihood = -3308.0473                             Pseudo R2     = 0.0103
```

F12G_läser_böcker_v	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
kon_egenid						
Tjej	.3918963	.0612168	6.40	0.000	.2719136	.5118789
Icke-binär/Osäker/Annan	1.129964	.2474781	4.57	0.000	.6449163	1.615012
utrikes						
Utrikes födda	.3632677	.0913071	3.98	0.000	.184309	.5422264
_cons	-1.161755	.0489172	-23.75	0.000	-1.257631	-1.065879

Tabell 21- Modell 2, exklusive observationer med partiellt bortfall

```
Iteration 0:   log likelihood = -3154.4401
Iteration 1:   log likelihood =  -3124.556
Iteration 2:   log likelihood = -3124.4051
Iteration 3:   log likelihood = -3124.4051

Logistic regression                                     Number of obs =  5,186
                                                        LR chi2(3)    =  60.07
                                                        Prob > chi2   = 0.0000
Log likelihood = -3124.4051                             Pseudo R2     = 0.0095
```

F12G_läser_böcker_v	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
kon_egenid						
Tjej	.3823638	.0631011	6.06	0.000	.2586878	.5060397
Icke-binär/Osäker/~n	1.081709	.2570435	4.21	0.000	.5779132	1.585505
utrikes						
Utrikes födda	.3482752	.0974428	3.57	0.000	.1572909	.5392595
_cons	-1.135446	.0505075	-22.48	0.000	-1.234439	-1.036453

Modell 3

Tabell 22. Modell 3, inklusive observationer med partiellt bortfall

```
Iteration 0:   log likelihood = -3342.6429
Iteration 1:   log likelihood = -3302.2461
Iteration 2:   log likelihood = -3302.0119
Iteration 3:   log likelihood = -3302.0119

Logistic regression                                     Number of obs =  5,525
                                                        LR chi2(7)    =  81.26
                                                        Prob > chi2   = 0.0000
Log likelihood = -3302.0119                             Pseudo R2     = 0.0122
```

F12G_läser_böcker_v	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
kon_egenid						
Tjej	.3932202	.0612911	6.42	0.000	.2730919	.5133486
Icke-binär/Osäker/Annan	1.137471	.2479742	4.59	0.000	.6514503	1.623491
utrikes						
Utrikes födda	.3305237	.0941835	3.51	0.000	.1459274	.5151201
Omradestyp2020						
2. Områden med socioekonomi..	.1630136	.1675061	0.97	0.330	-.1652924	.4913195
3. Socioekonomiskt blandade..	-.2265543	.1530376	-1.48	0.139	-.5265026	.073394
4. Områden med goda socioek..	-.1212158	.1445177	-0.84	0.402	-.4044654	.1620337
5. Områden med mycket goda ..	-.1881517	.1646759	-1.14	0.253	-.5109105	.1346072
_cons	-1.04028	.1451667	-7.17	0.000	-1.324801	-.7557583

Tabell 23. Modell 3, exklusive observationer med partiellt bortfall

```
Iteration 0:   log likelihood = -3154.4401
Iteration 1:   log likelihood = -3117.2184
Iteration 2:   log likelihood = -3117.0219
Iteration 3:   log likelihood = -3117.0219

Logistic regression                                     Number of obs =  5,186
                                                        LR chi2(7)    =  74.84
                                                        Prob > chi2   = 0.0000
Log likelihood = -3117.0219                             Pseudo R2     = 0.0119
```

F12G_läser_böcker_v	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
kon_egenid						
Tjej	.3851467	.0632059	6.09	0.000	.2612655	.5090279
Icke-binär/Osäker/~n	1.09675	.2575323	4.26	0.000	.5919956	1.601504
utrikes						
Utrikes födda	.3148677	.1002565	3.14	0.002	.1183686	.5113668
Omradestyp2020						
2. Områden med soc..	.2385813	.18059	1.32	0.186	-.1153687	.5925312
3. Socioekonomiskt..	-.2166369	.1659106	-1.31	0.192	-.5418157	.1085419
4. Områden med god..	-.1068942	.1573199	-0.68	0.497	-.4152356	.2014472
5. Områden med myc..	-.1702098	.1765904	-0.96	0.335	-.5163206	.1759009
_cons	-1.032465	.1583809	-6.52	0.000	-1.342886	-.7220441

Modell 4

Tabell 24. Modell 4, inklusive observationer med partiellt bortfall

```
Iteration 0:   log likelihood = -3342.6429
Iteration 1:   log likelihood = -3302.2461
Iteration 2:   log likelihood = -3302.0119
Iteration 3:   log likelihood = -3302.0119

Logistic regression                                     Number of obs =  5,525
                                                        LR chi2(7)    =  81.26
                                                        Prob > chi2   = 0.0000
Log likelihood = -3302.0119                             Pseudo R2     = 0.0122
```

F12G_läser_böcker_v	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
kon_egenid						
Tjej	.3932202	.0612911	6.42	0.000	.2730919	.5133486
Icke-binär/Osäker/Annan	1.137471	.2479742	4.59	0.000	.6514503	1.623491
utrikes						
Utrikes födda	.3305237	.0941835	3.51	0.000	.1459274	.5151201
1.Omradestyp2020_1	.1881517	.1646759	1.14	0.253	-.1346072	.5109105
1.Omradestyp2020_2	.3511652	.1317389	2.67	0.008	.0929617	.6093688
1.Omradestyp2020_3	-.0384026	.1115513	-0.34	0.731	-.2570392	.1802339
1.Omradestyp2020_4	.0669358	.097068	0.69	0.490	-.1233139	.2571856
_cons	-1.228431	.0965154	-12.73	0.000	-1.417598	-1.039265

Tabell 25. Modell 4, exklusive observationer med partiellt bortfall

```
Iteration 0:   log likelihood = -3154.4401
Iteration 1:   log likelihood = -3117.2184
Iteration 2:   log likelihood = -3117.0219
Iteration 3:   log likelihood = -3117.0219

Logistic regression                                     Number of obs =  5,186
                                                        LR chi2(7)    =  74.84
                                                        Prob > chi2   = 0.0000
Log likelihood = -3117.0219                             Pseudo R2     = 0.0119
```

F12G_läser_böcker_v	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
kon_egenid						
Tjej	.3851467	.0632059	6.09	0.000	.2612655	.5090279
Icke-binär/Osäker/~n	1.09675	.2575323	4.26	0.000	.5919956	1.601504
utrikes						
Utrikes födda	.3148677	.1002565	3.14	0.002	.1183686	.5113668
1.Omradestyp2020_1	.1702098	.1765904	0.96	0.335	-.1759009	.5163206
1.Omradestyp2020_2	.4087911	.1355286	3.02	0.003	.1431599	.6744222
1.Omradestyp2020_3	-.0464271	.1140075	-0.41	0.684	-.2698777	.1770235
1.Omradestyp2020_4	.0633156	.098781	0.64	0.522	-.1302917	.2569229
_cons	-1.202675	.0984029	-12.22	0.000	-1.395541	-1.009809

Modell 5

Tabell 26. Modell 5, inklusive observationer med partiellt bortfall

```
Iteration 0:   log likelihood = -3158.9824
Iteration 1:   log likelihood = -3111.3036
Iteration 2:   log likelihood = -3111.0314
Iteration 3:   log likelihood = -3111.0314

Logistic regression                                     Number of obs =  5,194
                                                        LR chi2(8)    =  95.90
                                                        Prob > chi2   = 0.0000
Log likelihood = -3111.0314                             Pseudo R2     = 0.0152
```

F12G_läser_böcker_v	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
kon_egenid						
Tjej	.3803083	.0632638	6.01	0.000	.2563136	.5043031
Icke-binär/Osäker/Annan	1.03792	.2580847	4.02	0.000	.532083	1.543756
utrikes						
Utrikes födda	.3081106	.1004401	3.07	0.002	.1112516	.5049697
1.Omradestyp2020_1	.2792836	.1785372	1.56	0.118	-.0706428	.62921
1.Omradestyp2020_2	.471989	.1364481	3.46	0.001	.2045557	.7394223
1.Omradestyp2020_3	.0285674	.1153098	0.25	0.804	-.1974357	.2545704
1.Omradestyp2020_4	.0966168	.0990139	0.98	0.329	-.0974468	.2906804
1.hogutb_vh	.3271742	.0709492	4.61	0.000	.1881163	.466232
_cons	-1.478398	.1162512	-12.72	0.000	-1.706246	-1.25055

Tabell 27. Modell 5, exklusive observationer med partiellt bortfall

```
Iteration 0:   log likelihood = -3154.4401
Iteration 1:   log likelihood = -3106.6998
Iteration 2:   log likelihood = -3106.4278
Iteration 3:   log likelihood = -3106.4278

Logistic regression                                     Number of obs =  5,186
                                                        LR chi2(8)    =  96.02
                                                        Prob > chi2   = 0.0000
Log likelihood = -3106.4278                             Pseudo R2     = 0.0152
```

F12G_läser_böcker_v	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
kon_egenid						
Tjej	.3838611	.0633296	6.06	0.000	.2597374	.5079848
Icke-binär/Osäker/~n	1.040695	.2580982	4.03	0.000	.5348319	1.546558
utrikes						
Utrikes födda	.3109187	.1004946	3.09	0.002	.113953	.5078844
1.Omradestyp2020_1	.2816872	.1786664	1.58	0.115	-.0684926	.6318669
1.Omradestyp2020_2	.475576	.1366014	3.48	0.000	.2078422	.7433098
1.Omradestyp2020_3	.0348168	.1154891	0.30	0.763	-.1915376	.2611712
1.Omradestyp2020_4	.1009351	.0992265	1.02	0.309	-.0935452	.2954155
1.hogutb_vh	.3230897	.0709888	4.55	0.000	.1839541	.4622253
_cons	-1.481853	.1165253	-12.72	0.000	-1.710238	-1.253467

Modell 6

```
Iteration 0:   log likelihood = -3154.4401
Iteration 1:   log likelihood = -3041.1364
Iteration 2:   log likelihood = -3039.6867
Iteration 3:   log likelihood = -3039.6855
Iteration 4:   log likelihood = -3039.6855

Logistic regression                                     Number of obs =  5,186
                                                        LR chi2(9)    = 229.51
                                                        Prob > chi2   = 0.0000
Log likelihood = -3039.6855                             Pseudo R2     = 0.0364
```

F12G_läser_böcker_v	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
kon_egenid						
Tjej	.3692128	.0642275	5.75	0.000	.2433292	.4950965
Icke-binär/Osäker/~n	.9854985	.2625141	3.75	0.000	.4709803	1.500017
utrikes						
Utrikes födda	.2271401	.103063	2.20	0.028	.0251404	.4291398
1.Omradestyp2020_1	.2489294	.1816997	1.37	0.171	-.1071954	.6050543
1.Omradestyp2020_2	.403146	.1392684	2.89	0.004	.1301849	.6761071
1.Omradestyp2020_3	.0236034	.1170368	0.20	0.840	-.2057845	.2529912
1.Omradestyp2020_4	.0920997	.1004126	0.92	0.359	-.1047055	.2889048
1.hogutb_vh	.3345161	.0721771	4.63	0.000	.1930516	.4759806
1.F11F_bibliotek_v	1.479813	.1312098	11.28	0.000	1.222647	1.73698
_cons	-1.547831	.1183139	-13.08	0.000	-1.779722	-1.31594

Tabell 28. Modell 6 med oddskvoter.

```
Logistic regression                                     Number of obs =  5,186
                                                        LR chi2(9)    = 229.51
                                                        Prob > chi2   = 0.0000
Log likelihood = -3039.6855                             Pseudo R2     = 0.0364
```

F12G_läser_böcker_v	Odds ratio	Std. err.	z	P>|z|	[95% conf.	interval]
kon_egenid						
Tjej	1.446595	.0929113	5.75	0.000	1.275488	1.640657
Icke-binär/Osäker/~n	2.679147	.7033138	3.75	0.000	1.601563	4.481764
utrikes						
Utrikes födda	1.255006	.1293446	2.20	0.028	1.025459	1.535936
1.Omradestyp2020_1	1.282652	.2330574	1.37	0.171	.8983501	1.831352
1.Omradestyp2020_2	1.496525	.2084187	2.89	0.004	1.139039	1.966209
1.Omradestyp2020_3	1.023884	.1198321	0.20	0.840	.8140085	1.287872
1.Omradestyp2020_4	1.096474	.1100998	0.92	0.359	.9005897	1.334965
1.hogutb_vh	1.397264	.1008504	4.63	0.000	1.212945	1.609592
1.F11F_bibliotek_v	4.392125	.5762898	11.28	0.000	3.396165	5.680161
_cons	.2127088	.0251664	-13.08	0.000	.168685	.2682221

Note: **_cons** estimates baseline odds.

.

Likelihood ratio-test

Tabell 29. Likelihood ratio-test för modell 1 och 2, inklusive observationer med partiellt bortfall

```
. lrtest modell10 modell20

Likelihood-ratio test
Assumption: modell10 nested within modell20

 LR chi2(1) =  15.39
Prob > chi2 = 0.0001
```

Tabell 30. Likelihood ratio-test för modell 1 och 2, exklusive observationer med partiellt bortfall

```
Likelihood-ratio test
Assumption: modell12 nested within modell22

 LR chi2(1) =  12.43
Prob > chi2 = 0.0004
```

Tabell 31. Likelihood ratio-test för modell 2 och ¾, inklusive observationer med partiellt bortfall

```
Likelihood-ratio test
Assumption: modell20 nested within modell40

 LR chi2(4) =  12.07
Prob > chi2 = 0.0168
```

Tabell 32. Likelihood ratio-test för modell 2 och 3/4, exklusive observationer med partiellt bortfall

```
Likelihood-ratio test
Assumption: modell22 nested within modell42

 LR chi2(4) =  14.77
Prob > chi2 = 0.0052
```

Tabell 33. Likelihood ratio-test för modell 4 och 5, inklusive observationer med partiellt bortfall

```
Likelihood-ratio test
Assumption: modell41 nested within modell51

 LR chi2(1) =  21.76
Prob > chi2 = 0.0000
```

Tabell 34. Likelihood ratio-test för modell 4 och 5, exklusive observationer med partiellt bortfall

```
Likelihood-ratio test
Assumption: modell42 nested within modell52

 LR chi2(1) =  21.19
Prob > chi2 = 0.0000
```

Tabell 35 Likelihood ratio-test för modell 5 och 6.

```
Likelihood-ratio test
Assumption: modell52 nested within modell62

 LR chi2(1) = 133.48
Prob > chi2 = 0.0000
```

AIC och BIC

Tabell 36. AIC och BIC för modell 1, inklusive observationer med partiellt bortfall.

Akaike's information criterion and Bayesian information criterion

Model	N	ll(null)	ll(model)	df	AIC	BIC
.	5,555	-3361.851	-3335.972	3	6677.943	6697.811

Note: BIC uses N = number of observations. See **[R] BIC note**.

Tabell 37. AIC och BIC för modell 1.

Akaike's information criterion and Bayesian information criterion

Model	N	ll(null)	ll(model)	df	AIC	BIC
.	5,186	-3154.44	-3130.622	3	6267.245	6286.906

Note: BIC uses N = number of observations. See **[R] BIC note**.

Tabell 38. AIC och BIC för modell 2, inklusive observationer med partiellt bortfall.

Akaike's information criterion and Bayesian information criterion

Model	N	ll(null)	ll(model)	df	AIC	BIC
.	5,525	-3342.643	-3308.047	4	6624.095	6650.563

Note: BIC uses N = number of observations. See **[R] BIC note**.

Tabell 39. AIC och BIC för modell 2.

Akaike's information criterion and Bayesian information criterion

Model	N	ll(null)	ll(model)	df	AIC	BIC
.	5,186	-3154.44	-3124.405	4	6256.81	6283.025

Note: BIC uses N = number of observations. See **[R] BIC note**.

Tabell 40. AIC och BIC för modell 4, inklusive observationer med partiellt bortfall.

Akaike's information criterion and Bayesian information criterion

Model	N	ll(null)	ll(model)	df	AIC	BIC
modell41	5,525	-3342.643	-3302.012	8	6620.024	6672.96

Note: BIC uses N = number of observations. See **[R] BIC note**.

Tabell 41. AIC och BIC för modell 4.

Akaike's information criterion and Bayesian information criterion

Model	N	ll(null)	ll(model)	df	AIC	BIC
.	5,186	-3154.44	-3117.022	8	6250.044	6302.473

Note: BIC uses N = number of observations. See **[R] BIC note**.

Tabell 42. AIC och BIC för modell 5, inklusive observationer med partiellt bortfall.

Akaike's information criterion and Bayesian information criterion

Model	N	ll(null)	ll(model)	df	AIC	BIC
.	5,194	-3158.982	-3111.031	9	6240.063	6299.06

Note: BIC uses N = number of observations. See **[R] BIC note**.

Tabell 43. AIC och BIC för modell 5.

Akaike's information criterion and Bayesian information criterion

Model	N	ll(null)	ll(model)	df	AIC	BIC
.	5,186	-3154.44	-3106.428	9	6230.856	6289.839

Note: BIC uses N = number of observations. See **[R] BIC note**.

Tabell 44. AIC och BIC för modell 6.

Akaike's information criterion and Bayesian information criterion

Model	N	ll(null)	ll(model)	df	AIC	BIC
.	5,186	-3154.44	-3039.685	10	6099.371	6164.908

Note: BIC uses N = number of observations. See **[R] BIC note**.

Cramer's V.

I detta avsnitt presenteras utdrag för Cramer's V och Pearson chi2-test från Stata.

Tabell 45. Cramers V och Pearson chi2 för kön och utrikes födda

```
          Pearson chi2(2) =   0.9486   Pr = 0.622
               Cramér's V =   0.0131
```

Tabell 46. Cramers V och Pearson chi2 för kön och områdestyp

```
          Pearson chi2(8) =   6.1997   Pr = 0.625
               Cramér's V =   0.0237
```

Tabell 47. Cramers V och Pearson chi2 för kön och föräldrars utbildning

```
          Pearson chi2(2) =  12.0572   Pr = 0.002
               Cramér's V =   0.0481
```

Tabell 48. Cramers V och Pearson chi2 för kön och biblioteksbesök

```
          Pearson chi2(2) =   8.1213   Pr = 0.017
               Cramér's V =   0.0383
```

Tabell 49. Cramers V och Pearson chi2 för utrikes födda och områdestyp

```
          Pearson chi2(4) = 297.4065   Pr = 0.000
               Cramér's V =   0.2315
```

Tabell 50. Cramers V och Pearson chi2 för utrikes födda och föräldrars utbildning

```
          Pearson chi2(1) =   2.6159   Pr = 0.106
               Cramér's V =  -0.0224
```

Tabell 51. Cramers V och Pearson chi2 för utrikes födda och biblioteksbesök

```
          Pearson chi2(1) =  40.3011   Pr = 0.000
               Cramér's V =   0.0852
```

Tabell 52. Cramers V och Pearson chi2 för områdestyp och föräldrars utbildning

```
          Pearson chi2(4) = 191.9523   Pr = 0.000
               Cramér's V =   0.1919
```

Tabell 53. Cramers V och Pearson chi2 för områdestyp och biblioteksbesök

```
          Pearson chi2(4) =  30.7148   Pr = 0.000
               Cramér's V =   0.0744
```

Tabell 54. Cramers V och Pearson chi2 för föräldrars utbildning och biblioteksbesök

```
          Pearson chi2(1) =   0.1931   Pr = 0.660
               Cramér's V =  -0.0061
```