Stockholms universitet

Faktorerna som får skolungdomar att ta en sup: En logistisk regressionsanalys

Emma Romero-Hamrin

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2023:7
Matematisk statistik
September 2023

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2023:7**
http://www.math.su.se/matstat

Faktorerna som får skolungdomar att ta en sup: En logistisk regressionsanalys

Emma Romero-Hamrin*

Juni 2023

Sammanfattning

Syftet med denna rapport är att undersöka faktorer som har starkast samband med en hög alkoholkonsumtion på helger hos skolungdomar. Detta görs genom en logistiska regressionsanalys av ett datamaterial från Portugal. Olika urvalsmetoder tillämpas för att utesluta faktorer för att ta fram en logistisk regressionsmodell som innehåller de variabler med starkast samband med responsvariabeln. Analysen som utförs visar att bland annat kön, hur mycket ungdomar går ut med vänner, familjesituationer samt om huruvida mamman till studenten är en hemma mamma eller ej är de faktorer med samband till hög alkoholkonsumtion.

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

Abstract

The purpose of this report is to investigate the factors most strongly associated with high weekend alcohol consumption among school adolescents. This is accomplished through a logistic regression analysis of a dataset from Portugal. Various selection methods are applied to exclude factors, in order to develop a logistic regression model containing the variables most strongly correlated with the response variable. The conducted analysis reveals that, among other factors, gender, frequency of social outings with friends, family situations, and whether the student's mother is a stay-at-home mom or not are correlated with high alcohol consumption.

Förord

Jag vill börja med att tacka mina handledare Daniel Ahlberg och Johannes Heiny som gett stöd, idéer och råd genom terminens gång.

Jag vill även tacka Abir Gharbi för stöd genom terminens gång och råd för ett förbättrat arbete.

Detta är ett kandidatexamensarbete som motsvarar 15 högskolepoäng i matematisk statistik som skrivits vid matematiska institutionen på Stockholms universitet.

Innehållsförteckning

1 Inledning

Alkohol har varit en del av samhället sedan lång tid tillbaka. Alkoholkonsumtion associeras många gånger till festligheter och njutning, men kan orsaka skada vid missbruk. Enligt forskning har det visat sig att gener har en betydande inverkan på risken att utveckla ett beroende av alkohol. Studier har uppskattat att gener kan svara för mellan 50 till 60 procent av variansen när det gäller benägenhet för alkoholberoende (Foroud, Edenberg och Crabbe 2010). Det betyder att resterande andel av variansen riskerar att utveckla ett alkoholberoende på grund utav riskfaktorer som exempelvis psykologiska skäl. Gränsen för lagligt drickande i Sverige är 18 år, men det stoppar inte ungdomarna. Enligt en rapport av Centralförbundet för alkohol- och narkotikaupplysning har 42% av nioendeklassare druckit alkohol under det senaste året och 69% av eleverna i gymnasiet årskurs två har druckit alkohol under det senaste året (Guttormsson 2020). Det är viktigt att få kännedom om varför en tonåring som än inte gått ur högstadiet vänder sig till en beroendeframkallande substans även om de känner till riskerna. Preventivt arbete för att minska risken för ett alkoholmissbruk i framtiden är oerhört viktigt.

Det finns indikationer på att personer med psykiska sjukdomar, trauma, historik av drogberoende eller påverkan av sociala faktorer, såsom att växa upp med en familjemedlem med alkoholberoende, har en ökad benägenhet att söka tröst eller lindring genom alkohol. Det är känsliga ämnen som inte alla vågar öppna upp sig om och därför är det viktigt att vara uppmärksam om man ser en indikation som kan tyda på risk för missbruk. Det är även viktigt att poängtera att drickande bland unga många gånger även kan bero på grupptryck där man söker en gemenskap med andra ungdomar som dricker.

I Portugal gjordes en enkät med frågeställningar om elevers alkoholvanor samt kring deras boende, föräldrars utbildningsnivå och arbete, fritidsaktiviteter, skolprestationer m.m. Denna enkät bidrar till att analysera vilka vardagliga faktorer som påverkar vissa elever till en hög alkoholkonsumtion. Resultatet visade att följande kriterier påverkar alkoholkonsumtionen mest: kön, social aktivitet, storlek på familj och familjerelationerna.

1.1 Syfte

Syftet med detta arbete är att identifiera de främsta riskfaktorerna bland olika faktorer i skolan. Detta ska göras genom att sätta upp en modell över hur konsumtion av alkohol beror på diverse faktorer. Datan kommer användas för att skatta parametrarna. Den slutgiltiga modellen ska även kunna prediktera hur stor risk det är att studenten har en hög alkoholkonsumtion på helgen, som presenteras i oddskvot som kan transformeras till en sannolikhet.

1.2 Data

Datamaterialet som valts är hämtad från Kaggle och heter Student Alcohol Consumption"(Cortez och Silva 2008) där datamängden med studenter som går i matematikklassen valdes. Datamängden är skapad via en undersökning i Portugal besvarad av 395 stycken elever och innehåller information om sociala och skolrelaterade faktorer. De variabler som finns är 30 stycken olika och av de kommer variabeln som representerar alkoholkonsumtion på helgen vara responsvariabeln. Det finns även en variabel som representerar daglig alkoholkonsumtion och båda dessa variabler har blivit besvarade på en skala 1-5, där 1 representerar väldigt låg alkoholkomsumtion och 5 representerar väldigt hög alkoholkonsumtion. Orsaken till att använda helgkonsumtion som responsvariabel istället för vardaglig konsumtion, trots att en vardaglig konsumtion bättre skulle beskriva en ohälsosam alkoholkonsumtion eller är mer intressant på grund av dess potentiella problematiska natur och sociala acceptans, är på grund av otillräcklig data. I Tabell 1 kan man se hur fördelningen är mellan de olika variablerna och vardaglig konsumtion har majoriteten av datapunkterna i de två lägsta nivåer med väldigt få datapunkter i de högre nivåerna. Alltså är helgkonsumtion en bättre lämpad responsvariabel eftersom den har en mer jämn fördelning av datapunkterna i de olika nivåerna. Detta leder till att variabeln som står för vardaglig konsumtion tas bort ur den statistiska analysen. En rapport från 2021 har dock visat att speciellt personer under 24 år, studenter och andra som dricker alkohol mer än två gånger i veckan anser att det är mer acceptabelt att dricka på vardagar än tidigare år vilket är ungefär åldersgruppen vi undersöker här (IQ 2021). I Figur 1 som ligger i appendix kan man se hur fördelningen av åldern är och man ser att främst 15 till 18 åringar har svarat på ankätet.

Konsumtion	1	2	3	4	5
Vardaglig	276	75	26	9	9
Helg	151	85	80	51	28

Tabell 1: Möjliga responsvariabler

Samtliga variabler med beskrivningar och typer har blivit samlade och presenteras i Tabell 2.

Variabelnamn	Beskrivning av variabel	Typ	Nivåer
Walc	Alkoholkonsumtion på helgen	Numerisk	1 (väldigt låg) - 5(väldigt hög)
school	Studentens skola	Binär	Gabriel Pereira/Mousinho da Silveira
sex	Studentens kön	Binär	Kvinna / Man
address	Studentens hemadress	Binär	Urban/Lantlig
famsize	Familjens storlek	Binär	LE3 (≤ 3)/GT3 (> 3)
Pstatus	Föräldrars boendesituation	Binär	Tillsammans/Isär
schoolsup	extra pedagogiskt stöd	Binär	Ja/Nej
famsup	familjens utbildningsstöd	Binär	Ja/Nej
paid	extra betalda klasser	Binär	Ja/Nej
activities	fritidsaktiviteter	Binär	Ja/Nej
nursery	Gick på förskola	Binär	Ja/Nej
higher	Vill ha högre utbildning	Binär	Ja/Nej
internet	Internetuppkoppling hemma	Binär	Ja/Nej
romantic	Med ett romantiskt förhållande	Binär	Ja/Nej
age	Studentens ålder	Numerisk	15-22
Medu	Mammans utbildning	Numerisk	0 (ingen) - 4 (högst)
Fedu	Pappans utbildning	Numerisk	0 (ingen) - 4 (högst)
traveltime	restid till skolan (minuter)	Numerisk	$1(< 15)$, $2(15-30)$, $3(30-60)$, $4(> 60)$
studytime	veckostudietid	Numerisk	$1(< 2h)$, $2(2-5h)$, $3(5-10h)$, $4(> 10h)$
failures	antal underkända klasser	Numerisk	n om $1 \leq 3$, annars 4
famrel	Kvaliteten på familjerelationerna	Numerisk	1 (väldigt dåligt) - 5 (excellent)
freetime	Fritid efter skolan	Numerisk	1 (väldigt lite) - 5 (väldigt mycket)
goout	Gå ut med vänner	Numerisk	1 (väldigt lite) - 5 (väldigt mycket)
health	Nuvarande hälsotillstånd	Numerisk	1 (väldigt dåligt) - 5 (väldigt bra)
absences	Antal skolfrånvaro	Numerisk	0-93
G1	Läsårets första betyget	Numerisk	0-20
G2	Läsårets andra betyg	Numerisk	0-20
G3	Slutbetyg	Numerisk	0-20
Mjob	Mammans jobb	Nominell	Lärare/hälsovård/civil/hemma/annat
Fjob	Pappans jobb	Nominell	Lärare/hälsovård/civil/hemma/annat
reason	anled. val av skola	Nominell	Nära hem/skolans rykte/kurs pref./annat
guardian	studentens vårdnadshavare	Nominell	Mamma/Pappa/Annan

Tabell 2: Variabler

I den statistiska analysen används endast en slumpmässigt utvald 70 procent mängd av datan då de resterande 30 procenten kommer användas för att testa modellernas prediktionsförmåga.

1.2.1 Dummy variabler

Datan som vi har innehåller både numeriska och kategoriska variabler som syns i Tabell 2. För att det ska vara lättare att arbeta med variablerna ska några variabler transformeras till så kallade dummy variabler. Dummy variabler är när man låter heltal representera en kategori, ett exempel som görs i denna rapport är att könets kategori transformeras så att kvinna representeras som talet 1 och man representeras som talet 0. Dummy variabler används

även för variabler med ordnade och nominella nivåer där man skapar $n-1$ nya variabler av en variabel med n nivåer (Hardy 1993). Alltså om vi tar Mjob från Tabell 2 som representerar mammans jobb som exempel som har fem nivåer som är lärare, hälsovård, civil, hemma och annat. Då skapar vi fyra nya variabler som skulle kunna vara som det syns i Tabell 3.

Orginal	Mjob_lärare	Mjob_hälsovård	Mjob_civil	Mjob_hemma
lärare	1	0	0	0
hälsovård	0	1	0	0
civil	0	0	1	0
hemma	0	0	0	1
annat	0	0	0	0

Tabell 3: Dummy transformation av Mjob

I Tabell 3 ser man att de fyra skapade dummy variablerna har fått tilldelade talet ett vid sin nivå och anledningen till varför det endast krävs fyra är för om alla dessa variabler skulle vara lika med noll skulle det betyda att mamman ej har någon av dessa yrken vilket betyder att hon tillhör nivån ännat". Denna typ av transformation görs på samtliga nominella och binära variabler. Hur resten av variablerna transformerades kan ses i Tabell 4 som ligger i Appendix A, men de viktiga variablerna att komma ihåg är de som presenteras i Tabell 5. För variabeln famsize står LE3 för mindre eller lika med tre i familjen och GT3 står för fler än tre i familjen.

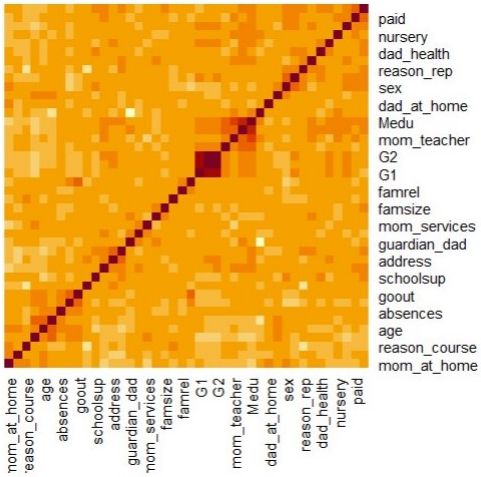
Variabel	Nivåer
sex	0-Man, 1-Kvinna
goout	1(väldigt lite) - 5(väldigt mycket)
mom_at_home	0 - Nej, 1 - Ja
famsize	0 - LE3, 1 - GT3
famrel	1(väldigt dålig) - 5(väldigt bra)
absences	0-93

Tabell 5: Viktiga variabler

1.2.2 Korrelationer

Korrelationer måste kontrolleras innan modellframtagandet påbörjas för att förstå relationerna mellan variablerna. Det finns tre korrelationstest vilket är Pearson, Spearman och Kendall. För detta dataset kommer vi använda Kendall's tau som korrelation vilket är ett icke-parametriskt mått som passar

då de numeriska variablerna *age* och *absences* som ej är normalfördelade. De finns även fler numeriska variabler som exempelvis *Medu* men de kan även hanteras som kategoriska variabler och resten av variablerna är binära. Kendall's tau kan anta värden mellan -1 och 1 där -1 tyder på ett negativt samband, 1 tyder på ett positivt samband medan 0 betyder att det inte finns något statistisk signifikans mellan variablerna. Ett korrelationsvärde på 0 betyder att det ej finns någon ordning i sambandet mellan variablerna men det betyder ej att det inte finns något linjärt samband mellan variablerna. Figur 2 visar en heatmap över variablernas korrelationer, ju mörkare färg desto högre korrelation är det mellan variablerna. Figuren visar en översiktlig bild över korrelationerna men sedan avläses de i en tabell för att se hur mycket korrelation det är.



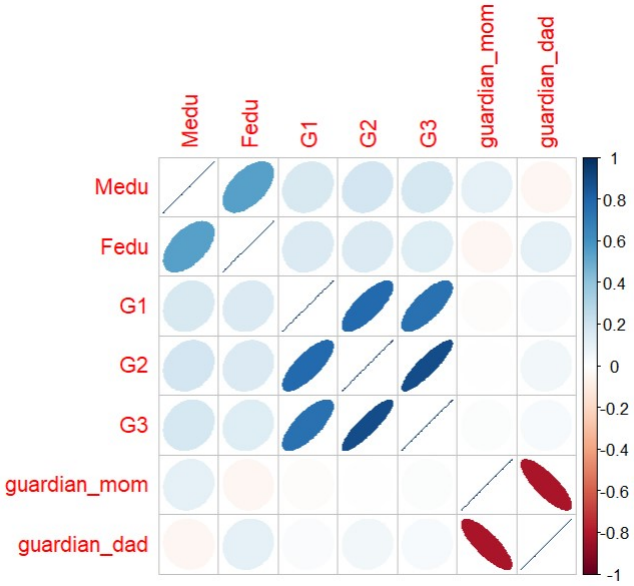
Figur 2: Heatmap av Kendall's tau

Alla korrelationer som var över 0.5 kontrollerades och de som hade den korrelationen var följande:

- Medu & Fedu: 0.549
- G1 & G2: 0.773
- G1 & G3: 0.746
- G2 & G3: 0.886
- guardian_mom & guardian_dad: -0.812

I Figur 3 kan man se hur de korrelationerna ovan ser ut grafiskt, det är alltså en illustration och ej en plot av data. Där ser man att G1, G2 och G3 är

korrelerade vilket betyder att det endast behövs en av de i modelleringen. De variablerna står för betyg över året och den variabeln som behålls är slutbetyget vilket är variabeln *G3*. Pågrund av korrelation tas även *Fedu* och *guardian_ dad* bort från möjliga kandidater för en potentiell modell.



Figur 3: Kendall's tau

1.3 Metod

Med hjälp av datan som vi har ska vi få fram en modell som visar de variablerna som har det största sambandet med responsvariabeln Walc. Denna modell ska även kunna prediktera hur stor risk man är att ha en hög alkoholkonsumtion i procent.

Responsvariabeln Walc är nu en ordnad variabel som går från ett till fem där ett är en väldigt låg och fem är en väldigt hög alkoholkonsumtion. För att ha den optimala variabeln för logistisk regression behöver responsvariabeln vara en binär variabel. Vi kommer att dela upp skalan 1-5 i två grupper med en låg och en hög grad. Alltså bestämmer vi oss för att dela upp analysen i tre delar. I följande analyser kommer responsvariabeln Walc transformeras så att en låg alkoholkonsumtion kommer sättas som noll och en hög alkoholkonsumtion kommer sättas som ett.

I första analysen kommer nivåerna 1 och 2 ses som låg alkoholkonsumtion medan nivåerna 3,4 och 5 ses som hög alkoholkonsumtion.

I andra analysen kommer nivåerna 1,2 och 3 ses som låg alkoholkonsumtion medan nivåerna 4 och 5 ses som hög alkoholkonsumtion.

I tredje och sista analysen kommer nivåerna 1,2,3 och 4 ses som låg alkoholkonsumtion medan endast nivå 5 anses som hög alkoholkonsumtion.

I dessa tre analyser kommer sedan en varsin modell tas fram med hjälp av metoderna Purposeful selection och olika stepwise selektioner. Dessa olika metoder kommer ge sina egna modeller och kommer testas genom att bland annat med likelihood ratio test, titta på dess VIF värden och AIC värden med mera. Efter testen kommer även modellerna kontrolleras i en så kallad ROC-kurva som kommer visa hur bra modellerna predikterar värden för Walc. Dessa metoder och tester kommer beskrivas i avsnitt 2 Teori. I slutet kommer vi alltså få en slutgiltig modell för varje analys och vi kan då se vilka förklarande variabler som modellerna fått. Därefter kan vi då jämföra vilka förklarande variabler som är gemensamma och vad som skiljer modellerna åt, vilket vi då kan reflektera över.

1.3.1 Programvara

Denna statiska analys kommer genomföras i programvaran R som används inom statistisk databehandling och datavetenskap. I detta program har främst biblioteket stats används för ‘glm()‘ funktionen vilket är till för modellframtagningen. Funktionen glm() används för generaliserade linjär modeller och används för att modellerna binär data där responsvariabeln endast kan anta två möjliga värden.

Biblioteket stats användes även för stepwise metoderna samt för att få AIC värdena för modellerna. Andra bibliotek som användes var lmtest för att göra likelihood ratio testen samt rcompanion för att räkna ut alla Pseudo R^2. Biblioteket ggplot2 användes för att plotta figurer och effect används för att visulisera effekterna av de förklarande variablerna i samband med responsvariabeln.

2 Teori

2.1 Logistisk regression

Regression är en metod som används för dataanalys när man vill beskriva en relation mellan en responsvariabel och en eller fler förklaringsvariabler. Logistisk regression är en metod för att skatta sannolikheten att en händelse ska inträffa baserat på de förklarande variablerna vilket används när responsvariabeln är binär. Denna modell är användbar i affärsbranschen där man exempelvis räknar ut sannolikheten att någon betalar sin räkning i tid vilket

görs genom att använda sig av parametrar som lön, skulder, hur stor räkningen är och så vidare. Detta är användbart för affärsbranschen för att se om det är värt att sälja till denna kund. (Agresti 2002) Logistisk regression är användbart i fler fält såsom medicinsk forskning, marknadsföring, försäkring och det som denna rapport är baserad på, samhällsvetenskap där man gör undersökningar och räknar ut sannolikheten för speciella beteenden beroende på olika faktorer ur undersökningen.

I en logistisk regression har vi en responsvariabel Y som måste vara binär och de förklarande variablerna X vilket ger att $\pi(x)$ som räknar ut sannolikheten att Y är lika med 1 givet x, detta kan uttryckas som $\pi(x) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x)$. Den logistiska modellen skrivs

$$\pi(x) = \frac{exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)} \tag{1}$$

och där oddskvoten vilket också kallas för logit modell räknas ut med ekvationen:

$$logit[\pi(x)] = ln(\frac{\pi(x)}{1 - \pi(x)}) = \beta_0 + \beta_1 x. \tag{2}$$

För att få en komplett modell måste värdena på de okända parametrarna β_0 och β_1 hittas. Dem okända parametrarna uppskattas genom att använda metoden maximum likelihood. Maximum likelihood uppskattar parametrar som maximerar sannolikhetsfunktionen (likelihooden) av datan vilket betyder att att vi försöker hitta de värdena för parametrarna som ger högst sannolikhet för det observerade fallet. För att räkna ut maximum likelihood måste först en sannolikhetsfunktion skapas (likelihood funktion) som får fram en datans sannolikhet av de okända parametrarna. Maximum likelihood är då de parametrar som maximerar sannolikhetsfunktionen och de parametrarna är de som förklarar datan bäst. (Hosmer Jr, Lemeshow och Sturdivant 2013)

Tidigare beskrev vi hur $\pi(x)$ räknar ut sannolikheten av att responsvariabeln är lika med 1, detta ger att sannolikheten att responsvariabeln är lika med noll är alltså $1 - \pi(x)$. Alltså vid för (x_i, y_i) där $y_i = 1$ har man sannnolikhetsfunktionen $\pi(x_i)$ och när $y_i = 0$ har man sannolikhetsfunktionen $1 - \pi(x_i)$. Då observationerna anses vara oberoende kan paret användas för sannolikhetsfunktionen (likelihood funktionen) så att:

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}. \tag{3}$$

Matematiskt är det lättare att räkna ut med log-likelihooden vilket är:

$$L(\beta) = ln[l(\beta)] = \sum_{i=1}^{n} y_i ln[\pi(x_i)] + (1 - y_i) ln[1 - \pi(x_i)]. \tag{4}$$

För att hitta det beta i Ekvation 4 som maximerar $L(\beta)$ delas ekvationen upp hänsyn till de sökta parametrarna β_0 och β_1:

$$\frac{\partial L(\beta)}{\partial \beta_0} = \sum_{i=1}^{n}(y_i - \pi(x_i)) = 0, \tag{5}$$

$$\frac{\partial L(\beta)}{\partial \beta_1} = \sum_{i=1}^{n} x_i(y_i - \pi(x_i)) = 0, \tag{6}$$

för alla i mellan 1 och n. Den β som fås ut av Ekvation 5 och 6 är maximum likelihood skattade och skrivs som $\hat{\beta}$. Ekvation 5 ger alltså att $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(x_i)$ vilket står för att antalet observerade händelser ska vara lika med antalet förutsagda händelser. Ekvation 6 tar hänsyn till den förklarande variabeln x_i där summan över skillnaderna mellan den observerade responsvariabeln y_i och den förutsagda sannolikheten $\pi(x_i)$ för varje observation ska vara lika med noll. (Hosmer Jr, Lemeshow och Sturdivant 2013)

Vi har nu gått igenom logistisk regression för endast en variabel men i denna rapport kommer en multipel logistisk regression genomföras eftersom datasett vi har innehåller mer än en variabel. Har man ett datasett med p stycken förklarande variabler $x_1, x_2, ..., x_p$ blir sannolikheten uttryck som $P(Y = 1|x) = \pi(x)$ och logit modellen av den multipla regressionen blir

$$g(x) = logit(\pi(x)) = ln(\frac{\pi(x)}{1 - \pi(x)}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p, \tag{7}$$

alltså blir den multipla logistiska regressionsmodellen för ett lyckat utfall som är en sigmoidfunktion:

$$\pi(x) = P(Y = 1|X = x) = \frac{e^{g(x)}}{1 + e^{g(x)}}. \tag{8}$$

För att estimera värdena för samtliga beta i den multivariata modellen så maximerar man log-likelihood funktionen. Detta hittas genom att lösa:

$$\frac{\partial L(\beta)}{\partial \beta_j} = \sum_{i=1}^{n}[y_i - \pi(x_i)]x_{ij} = 0, \tag{9}$$

där x_{ij} är värdet på den förklarande variabeln j för observation i. Detta ger ekvationssystemet

$$\sum_{i=1}^{n}[y_i - \pi(x_i)] = 0, \tag{10}$$

$$\sum_{i=1}^{n} x_{ij}[y_i - \pi(x_i)] = 0, \tag{11}$$

för alla $j = 1, 2, ..., p$. (Hosmer Jr, Lemeshow och Sturdivant 2013) Detta ekvationssytem löses med en optimeringsmetod som till exempel Newton-Raphson eller en gradientbaserad metod. Ekvation 10 betyder att antalet observerade händelser ska vara lika med antalet förutsagda händelser alltså ska skillnaden mellan responsvariabeln y_i och den förutsagda sannolikheten $\pi(x_i)$ för varje observation i vara lika med noll. Ekvation 11 är specificerad för den förklarande variabeln x_{ij} och där summan över skillnaderna mellan den observerade responsvariabeln och den förutsagda sannolikheten för varje observation ska vara lika med noll.

Beta koefficienterna man får ur maximum likelihood metoden används för logit modellen som visas i Ekvation 7. Ett $\beta_i > 0$ betyder att den parametern ökar sannolikheten att få ett lyckat fall om x_i ökar, ett $\beta_i < 0$ betyder att parametern minskar sannolikheten att få ett lyckat fall om x_i ökar och ett $\beta_i = 0$ betyder att parametern ej gör någon skillnad på utfallet.

2.2 VIF

Variansinflationsfaktor som förkortas till VIF är ett mått som kollar multicollinearitet alltså efter en hög korrelation mellan de förklarande variablerna i en regressionsmodell. Ett högt VIF värde tyder på mycket multicollinearitet vilket kan tyda på att modellen är dålig och opålitlig vilket i regressionsanalys kan ge höga standardfel samt höga p värden. Om man har $p - 1$ förklarande variabler får man

$$VIF_i = \frac{1}{1 - r_i^2}, i = 1, ..., p - 1, \tag{12}$$

där r_i^2 är determinationskoefficienten vilket beräknas som kvadraten av korrelationskoefficienten mellan reponsvariabeln och den förklarande variabeln som testas. Determinationskoefficienten är alltid mellan 0 och 1, ju närmare värdet är ett desto bättre passar modellen datan.

Ett allmänt kriterium för när ett VIF värde är för högt ligger oftast då $VIF \geq 5$ eller då $VIF \geq 10$. (Craney och Surles 2002)

2.3 AIC

AIC står för *Akaike information criterion* och den utvärderar hur bra en modells anpassade värden är jämfört med datans verkliga medelvärden. Den används främst för att jämföra olika modeller för att se vilken av flera modeller passar datan bäst dock säger AIC inget om hur bra en modell predikterar framtida värden. Formeln för AIC är

$$AIC = -2(\hat{L} - k) \tag{13}$$

där $\hat{L}$ står för maximum log likelihood och k står för antal parametrar i modellen. De modeller med låga AIC värden är de som anses vara de bäst passande modellerna (Agresti 2013).

2.4 Likelihood ratio test

Ett test för multinomiala modeller är ett likelihood ratio test som testar om skillnader mellan framtagna modeller från samma datasett. Man testar exempelvis en modell som innehåller samtliga variabler från en datasett mot en ny modell som blivit framtagen av någon statistisk metod för att se om det blivit en signifikant förbättring. Testet uttrycks i ett hypotestest där man oftast har att nollhypotesen är att den nya modellen ej gör något skillnad och att mothypotesen är att den nya modellen är signifikant bättre än den gamla. Testet räknar ut skillnaden på modellerna likelihood värden och betecknas som G^2 där formeln är

$$G^2 = -2log(\frac{L_0}{L_1}) \tag{14}$$

där L_0 och L_1 är maximum likelihood funktionerna för modellerna. Modellen L_0 står för nullmodellen vilket är den mindre modellen och L_1 är den modell med fler parametrar. Hypotestestet följer en χ^2 fördelning. Om man gör ett likelihood ratio test och får ett p värde som är mindre än 0.05 betyder det att den större modellen (L_1) är ett bättre val medans om man får ett p värde som är över 0.05 betyder de att båda passar datan lika bra men då tar man den mindre modellen (L_0) då de variablerna som modellerna inte har gemensamt ej ger något signifikant skillnad (Agresti 2013).

2.5 Purposeful selection

När man ska göra en modell ur ett datasett med många ko-variater är det svårt att veta vilka som ska vara med och inte för att få den bästa modellen. En av metoderna som ska användas är *purposeful selection* som är en metod med sju steg.

Steg 1:
I första steget ska alla förklarande variabler göra en varsin envariabelsanalys där de variabler med ett p värde högre än 0.25 vals bort som möjliga variabler för våran multivariata modell. Anledningen till att p värdes gränsen är mycket högre än den vanliga 0.05 gränsen är på grund av andra arbeten visat att en gräns på 0.05 kan ta bort variabler som egentligen är viktiga.
Steg 2:
Alla variabler som hade kriteriet i steg 1 sätts nu in i en multivariat modell. De förklarande variablerna som har ett p värde i Wald statistikan som är högre än 0.05 tas bort. När man kollat alla variablers p värden så ska den nya mindre modellen med endast de signifikanta variablerna testas mot modellen med samtliga variabler med ett likelihood ratio test. Detta visar då om den nya är bättre.
Steg 3:
I detta steg ska vi nu jämföra värdena i den nya mindre modellen med den större originella modellens värden. Om något värde skulle ha att $\Delta\hat{\beta} > 20\%$ betyder det att någon av de variablerna som togs bort ur den originella modellen är viktig och behöver läggas in i den nya modellen. Uttrycket $\Delta\hat{\beta}$ står för förändringen i koefficienten mellan nya mindre modellen och den större originella modellen. Alltså måste man cirkulera mellan steg 2 och 3 tills man hittar en ny modell där alla variabler är signifikanta och ingen viktig variabel har blivit exkluderat.
Steg 4:
Nu ska alla exkluderade variabler från steg 1 läggas in en i taget till modellen från steg 2 och 3 för att se om de variablerna är signifikanta för modellen, vilket kan göras med att kolla på Wald statistikans p värde. Detta görs för att kolla om de finns variabler som ej signifikanta till resultatet men blir signifikanta när de är i samband med andra variabler. Modellen som man får efter detta steg kallas *den preliminära huvudeffekts modellen.*

Steg 5:
När vi nu har *den preliminära huvudeffekts modellen* ska vi kolla närmare på alla variabler i modellen. För alla kontinuerliga variabler måste man kolla så att de är linjära med logiten. Detta betyder att man plottar variabeln mot logit-transformationen av responsvariabeln och kontrollera att det finns ett linjärt samband. Efter detta steg kallas modellen för *huvudeffekts modellen.*
Steg 6:
I detta steg ska man nu kolla efter samspel mellan variabler där man väljer vilka samband man ska ha med både från en statistisk och praktiska perspektiv. Detta betyder att de samband man väljer att lägga in i modellen måste vara rimliga ur ett kliniskt perspektiv. Man gör upp sin lista med möjliga samband och lägger in de en och en i *huvudeffekts modellen* och ser om de ger en signifikant skillnad med ett likelihood ratio test där signifikansgränsen är satt på 5% eftersom ett samband i en modell som ej är signifikant ökar standardfelen utan att effekterna av värdena ökar. Modellen vi får efter detta steg är låst och kallas för *den preliminära slutgiltiga modellen.*
Steg 7:
Innan *den preliminära slutgiltiga modellen* kan bli den slutgiltiga modellen måste den analyseras för att se om den passar. Detta steg utförs i alla metoder när man tar fram en modell.(Hosmer Jr, Lemeshow och Sturdivant 2013)

2.6 Stepwise metoder

En stepwise metod tar fram den bästa modellen genom att antingen addera eller ta bort en variabel i taget baserat på ett kriterium. I denna rapport användes en step funktion som använde sig av Akaikes Informations Criterium som förkortas som AIC (Ruengvirayudh och Brooks 2016).

2.6.1 Forward

I en forward selektion börjar man med en modell utan några variabler, därefter i varje steg så välj den variabel som ger den bästa modellen. Man fortsätter addera variabler tills att ingen variabel längre gör någon signifikant skillnad för modellen. I denna analys används AIC vilket betyder att selektionen kommer avslutas när AIC värdet slutar minska då ett lågt AIC värde indikerar på en bra modell (Ruengvirayudh och Brooks 2016).

2.6.2 Backward

I backward elimination börjar man med en modell som har alla ens variabler. Sedan tar man bort en variabel i taget som är minst signifikant vilket i detta fall är den variabel med högt AIC. Eliminationen avslutas när man ser att borttagning av variabler ej längre förbättrar modellen. (Ruengvirayudh och Brooks 2016)

2.6.3 Both

Denna typ av stepwise regression kombinerar forward selektion och backward elimination genom att addera variabler en i taget medans variabler även tas bort en i taget. Även i denna välj den modell som har bäst AIC (Hastie och Pregibon 1992).

2.7 Pseudo R squared

I linjär regression används R^2 som ett passforms test för modeller men det finns inte ett sådant för logistisk regression. Pseudo R^2 är ett liknande test för modellens förklaringsgrad. Eftersom det är ett liknande test testas tre olika R^2 då det ej finns en allmän formel men en allmän fakta är att ett högt R^2 tyder på en bra passad modell. Dessa index värden hamnar endast mellan 0 och 1. Pseudo R squared har ingen gräns som säger att detta värde tyder på en bra modell, utan de används istället för att jämföra modeller med varandra där ett högre Pseudo R squared är bättre (Walker och Smith 2016). I följande formler står *Full* för en modell med fler variabler och *Null* står för en modell med färre variabler.

McFadden:
Denna är den enklaste och vanligaste metoden som jämför log-likelihooden av de modeller som jämförs. Formeln är

$$R^2_{MF} = 1 - \frac{LL(Full)}{LL(Null)}. \tag{15}$$

Cox and Snell (ML):
Formeln för Cox and Snells R^2 är

$$R^2_{CS} = 1 - (\frac{L(Null)}{L(Full)})^{2/N} \tag{16}$$

där $L(Null)$ och $L(Full)$ är likelihoodfunktioner och N står för antalet observationer i data settet.

Nagelkerke (Cragg and Uhler):
Denna formel är en variant av Cox and Snells och görs genom att ta kvoten av dess index med dess max värde. Formeln är

$$R_N^2 = \frac{1 - (L(Null)/L(Full))^{2/N}}{1 - L(Null)^{2/N}}. \tag{17}$$

2.8 ROC AUC

Receiver operator characteristic är förkortat som ROC vilket är en grafisk kurva som visar hur bra en modell predikterar riktiga värden. Kurvan visar hur korrekt modellen predikterar ett positivt exempel (vilket i denna analys är hög alkoholkonsumtion) och hur bra den predikterar ett negativt exempel (vilket i denna är en låg alkoholkonsumtion). Punkterna för kurvan tas fram genom att räkna ut sannolikheterna för rätt predikterat vilket görs genom att beräkna andelen falska positiva och sanna positiva och jämföra modellens skattning med exempel från testdata. På x-axeln kommer man ha False positive rate"vilket är antalet felaktigt klassade negativa exempel dividerat med totala negativa exempel. Illustrerat som en formel blir det:

$$\text{False positive rate } = \frac{\text{Falska positiva}}{\text{Falska positiva} + \text{Sanna negativa}}.$$

På y-axeln kommer man ha True positive rate"vilket är antalet korrekt klassade positiva exempel dividerat med totala positiva exempel. Illustrerat som en formel blir det:

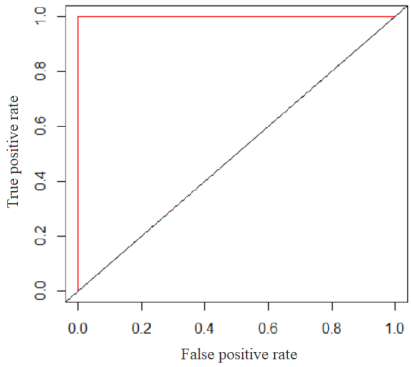
$$\text{True positive rate} = \frac{\text{Sanna positiva}}{\text{Sanna positiva} + \text{Falska negativa}}.$$

Dessa värden räknas ut för flera tröskelvärden och detta ger de punkter som ger en ROC kurva. Tröskelvärde är den gräns som avgör vilket klass ett exempel tillhör i en binär klassificering. Till exempel om modellen ger ett en sannolikhet på 0.6 att studenten har en hög alkoholkonsumtion och tröskelvärdet är satt på 0.5 betyder det att studenten är klassad som en hög alkoholkonsumerare men om tröskelvärdet skulle vara på 0.7 skulle 0.6 betyda att studenten har låg alkoholkonsumtion.

Hur bra modellen kan prediktera värden kan sammanfattas i ett så kallat AUC vilket står för ärea under the curve"vilket betyder area under kurvan. AUC värdet kan vara mellan 0.5 till 1 där ett högre värde visat på bättre prediktionsförmåga (Muschelli III 2020). Ett AUC värde på 0.5 anser att modellen är meningslös, ett AUC mellan 0.7-0.8 betyder att modellen är acceptabel, ett AUC mellan 0.8-0.9 betyder att modellen är mycket bra och

ett AUC värde över 0.9 anser att modellen är perfekt (Hosmer Jr, Lemeshow och Sturdivant 2013).

En illustration av som visar en ROC-kurva av en modell som predikterar perfekt kan synas i Figur 4.



Figur 4: ROC

3 Statistisk analys

3.1 Modell framtagning och testing

3.1.1 Responsvariabel 1

I denna modell framtagning kommer responsvariabeln vara satt att 1-2 är satt som låg alkoholkonsumtion och 3-5 är satt som hög alkoholkonsumtion. Fördelningen av reponsvariabeln är 236 stycken med låg alkoholkonsumtion och 159 stycken punkter med hög alkoholkonsumtion. Detta dataset delas in i två delar med ett dataset som innehåller 70 procent och en med 30 procent, datasetet med 70 procent av den beskrivna datan kommer användas för modelleringen som nu kommer beskrivas hur den genomfördes.

Purposeful selection:

Steg 1: De variablerna som fick ett p värde lägre än 0.25 och därav behölls blev school, sex, age, famsize, Pstatus, studytime, famsup, famrel, freetime, goout, health, absences, mom_health, dad_teacher, dad_services, dad_at_home.

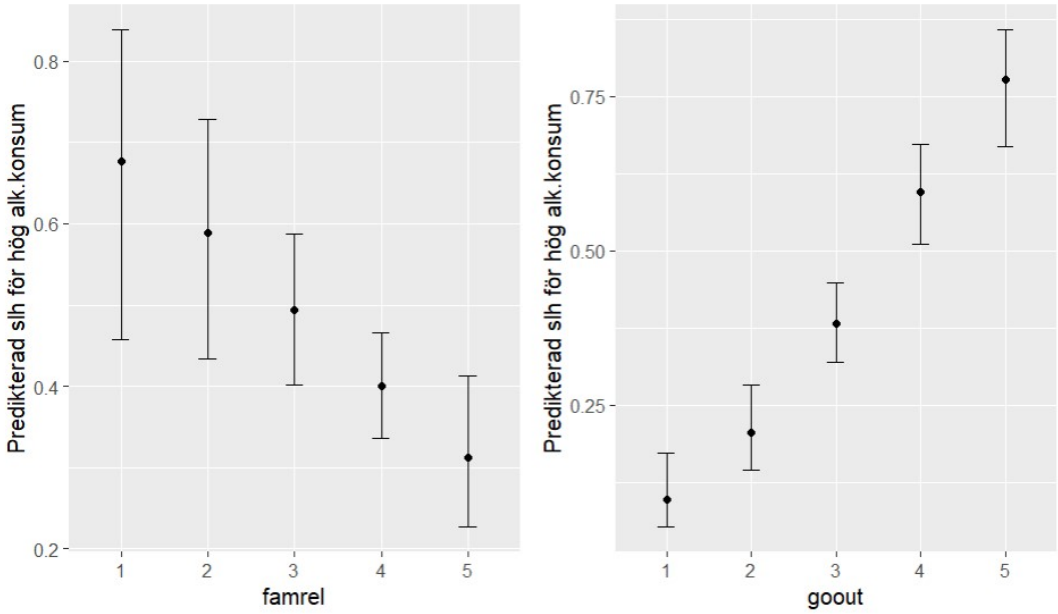
Steg 2: Nu när variablerna i steg 1 sätts in en multivariatmodell tas alla de variablerna som har ett p-värde högre än 0.05 bort. De variablerna som hade ett p-värde lägre än 0.05 sätts in i en

ny modell och dessa blev famrel, goout och dad_at_home. Modellen med variablerna från steg 1 och modellen med variablerna från detta steg testas nu mot varandra för att se om borttagningen av variabler i steg 2 ger en mer signifikant modell eller ej. Som beskrivet i avsnitt 2.5 används ett likelihood ratio test där nollhypotesen är att den större och mindre modellen passar lika bra och den alternativa hypotesen är att den större modellen är bättre. Loglikelihood värdet för den större modellen blev -143.89 och den mindre modellen fick ett värde på -159.06 med frihetsgrader på 17 respektive 4. Vilket gör man kan räkna ut Chi värdet med hjälp av Ekvation 14 men vi fick detta värde av funktionen som användes i R vilket blev 30.333. P-värdet för en chi-fördelning med en differens på 13 frihetsgrader ger ett p-värde på 0.004216 vilket betyder att det är mindre är gränsen 0.05. Detta betyder att man kan förkasta nollhypotesen vilket betyder i detta fall att den större modellen passar bättre.

Steg 3: Eftersom den större modellen är en bättre passning enligt likelihood ratio testet så kollar vi extra noga i detta steg. Om något av variablerna har ett $\Delta\beta$ som är större än 20 procent betyder det att någon av de variablerna som togs bort i steg 2 var viktig. Förändringen av koefficienten i detta fall var mindre än 0.20 i samtliga variabler vilket tyder på att inga viktiga variabler saknas. Med dessa resultat kommer vi fortsätta med den mindre modellen eftersom den större innehåller variabler med p-värden över 0.05.

Steg 4: För att se till att inga signifikanta variabler tagits bort i steg 1 läggs dessa variabler en i tagen in i den aktuella modellen. Detta gjordes genom likelihood ratio test där samtliga sa att den aktuella modellen var bättre. Undersökte även p-värdet för den adderade variabeln för att de om den är signifikant och detta var aldrig fallet.

Steg 5: Vi kontrollerade sedan att modellens variabler är linjära med logiten. För variabeln dad_at_home så behövs det ej kontrolleras eftersom den är binär, alltså antas det att den är linjär. För famrel samt goout så kollades linjäriteten genom att räkna ut oddskvoterna och i Figur 5 ser man att båda dessa är linjära så de kan behållas i modellen.



Figur 5: Logit-Transformationsplot

Steg 6: Alla möjliga samband undersöktes men ingen av de gav en signifikant skillnad.

Steg 7: Dessa steg kommer genomföras längre ner i denna analys i samband med jämförlig av en annan modell.

Efter att ha följt metoden fick man en modell som kallas för ModellP1 med variblerna *famrel, goout, dad_at_home.*

Stepwise:
Metoderna both och backward gav identiska modeller så denna modell kommer i framtiden kallas Backwards1. Backwards1 fick en modell med variablerna *sex, famsize, Pstatus, studytime, activities, famrel, freetime, goout, absences, dad_teacher, dad_at_home.* Forward metoden kommer ej fortsätta analyseras då den får fram en modell med väldigt många variabler och med höga p-värden. Samtliga stepwise modeller testades även med samspel men dessa blev bortselekterade i metoden.

Dessa modeller ska nu testas för att se vilken av de som är bäst passade för datan. Båda modellerna har bra VIF värden där ingen av de överstiger fem. Övriga test som utfördes på modellerna presenteras i Tabell 6,7 och 8. I Tabell 6 utförs ett likelihood ratio test och där testar man modellen

mot en modell utan några förklarande variabler. Den modellen Walc∼1 har 1 frihetsgrad (df) samt en loglikelihood värde på -188.96.

Tabell 6 redovisar att båda modellerna passar bättre än modellen utan några förklarande variabler eftersom p-värdet är mindre än 0.05 vilket betyder att man kan förkasta nollhypotesen som är att den större och mindre modellen passar datan lika bra. Eftersom nollhypotesen förkastas så behåller man den större modellen. Detta tyder på att de förklarande variablerna i ModellP1 samt Backwards1 gör en signifikant skillnad i förklarande syfte och är relevanta.

Modell	df	LogLik	χ^2	p-värde
ModellP1	4	-159.06	59.796	$6.5 \cdot e^{-13}$
Backwards1	12	-144.79	88.338	$3.5 \cdot e^{-14}$

Tabell 6: Likelihood ratio test

Ett ytterligare likelihood ratio test genomfördes mellan modellerna vilket kan ses i Tabell 7. Där redovisas ett p-värde som är mindre än 0.05 vilket betyder att även här förkastas nollhypotesen vilket betyder att den större modellen är en bättre passning vilket är Backwards1.

Modell	Δdf	χ^2	p-värde
ModellP1/Backwards1	8	28.543	0.0003813

Tabell 7: Likelihood ratio test mellan modellerna

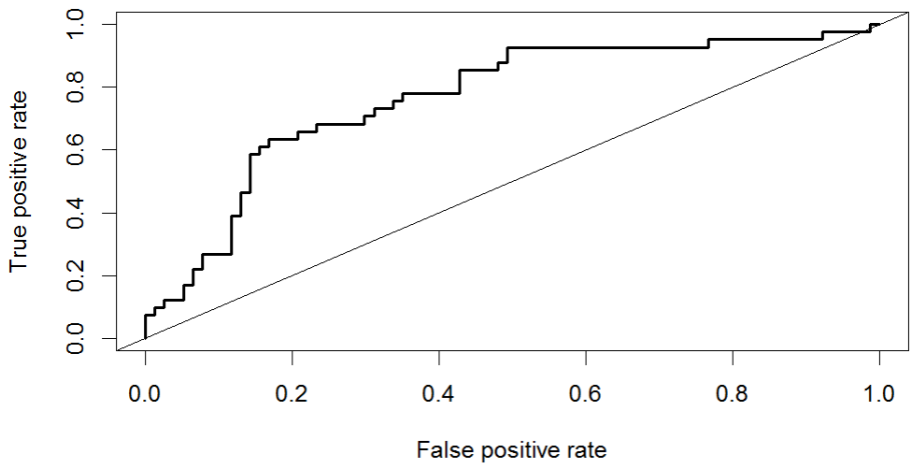
Tabell 8 kollar man värdena av Pseudo R^2 samt AIC värdena. Man ser att Backwards1 har ett lägre AIC samt högre R^2 värden vilket tyder på att den är en bättre passad modell än ModellP1.

Modell	AIC	McFadden R^2	Cox&Snell R^2	Nagelkerke R^2
ModellP1	326.12	0.1582	0.1942	0.2608
Backwards1	313.57	0.2338	0.2731	0.3668

Tabell 8: AIC och R^2

Efter att studerat de olika testen i Tabell 6-8 så visade det sig att Backwards1 var den mest lämpade modellen.

Vi kollade även dess ROC-kurva och där var även Backwards1 den bästa modellen med ett AUC värde på 0.766 medans ModellP1 hade ett AUC värde på 0.686. Backwards1:s ROC-kurva kan man se i Figur 6. Datan som testade modellerna alltså för ROC-kurvan är den andra delen av data settet som beskrivs i början av analysen, alltså 30 procent av den totala datan.



Figur 6: ROC-kurva för Backwards1

Modellen för responsvariabel 1 är sammanfattad i Tabell 9 med dess skattade värden och standardfel.

Koefficienter	Skattade värde	Standardfel	p-värde
Intercept	-1.64	1.02	0.109
sex	-0.58	0.32	0.065
famsize	-0.53	0.33	0.109
Pstatus	1.05	0.55	0.057
studytime	-0.36	0.20	0.076
activities	-0.43	0.30	0.149
famrel	-0.43	0.17	0.011
freetime	0.26	0.16	0.104
goout	0.88	0.15	$7.97 \cdot e^{-9}$
absences	0.027	0.02	0.145
dad_teacher	-1.05	0.64	0.097
dad_at_home	-1.63	0.74	0.028

Tabell 9: Slutgiltiga modellen för responsvariabel 1

Modellen utskriven syns i Ekvation 18 där variabel *sex* står för kvinna när $x_{sex} = 1$ och man när $x_{sex} = 0$. Vi har även *famsize* som står för familjen storlek där familjen är mindre eller lika med tre när $x_{famsize} = 0$ och familjen är fler än tre när $x_{famsize} = 1$. En annan binär variabel är *Pstatus* där $x_{Pstatus} = 0$ står för att föräldrarna är isär och $x_{Pstatus} = 1$ står för att föräldrarna är tillsammans. Variabeln *studytime* är numerisk och står för hur mycket eleven studerar där $x_{studytime} = 1$ står för mindre än två timmars plugg, $x_{studytime} = 2$ står för mellan 2-5 timmars plugg, $x_{studytime} = 3$ står för 5-10 timmars plugg och $x_{studytime} = 4$ står för mer än 10 timmars plugg i veckan. Variabeln *famrel* står för kvaliteten på familjerelationer och är numerisk där x_{famrel} kan vara mellan 1-5 som står för väldigt dålig familjerelation till väldigt bra familjerelation och *freetime* står för hur mycket fritid eleven har och är mellan 1-5 som står väldigt lite till väldigt mycket fritid. Variabeln *goout* står för hur mycket studenten går ut med vänner och där kan x_{goout} vara mellan 1-5 som står för ute med vänner väldigt lite till väldigt mycket, *absences* står för antal frånvaro och där kan $x_{absences}$ vara mellan 0-93. Resterande variabler är binära som kan vara antingen $x = 1$ vilket står för ja och $x = 0$ står för nej.

$$\begin{aligned} logit(\pi(x)) = -1.64 - 0.58x_{sex} - 0.53x_{famsize} + 1.05x_{Pstatus} - 0.36x_{studytime} \\ -0.43x_{activities} - 0.43x_{famrel} + 0.26x_{freetime} + 0.88x_{goout} + 0.027x_{absences} \\ -1.05x_{dad\_teacher} - 1.63x_{dad\_at\_home} \end{aligned} \tag{18}$$

3.1.2 Responsvariabel 2

I denna modell framtagning kommer responsvariabeln vara satt att 1-3 är satt som låg alkoholkonsumption och 4-5 är satt som hög alkoholkonsumtion. Fördelningen på variablerna är 316 stycken studenter rapporterade med låg alkoholkonsumtion och 79 stycken med hög alkoholkonsumtion. Detta datasett delas in i två delar med ett dataset som innehåller 70 procent och ett med 30 procent, den datasettdelen med 70 procent av den beskrivna datan kommer användas för modelleringen som nu kommer beskrivas hur den genomfördes.

Purposeful selection:

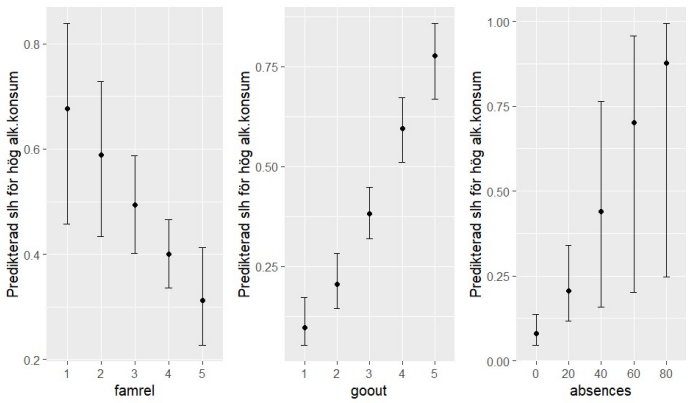
Steg 1: De variablerna som fick ett p-värde under 0.25 var sex, famsize, traveltime, studytime, failures, famsup, higher, famrel, freetime, goout, health, absences, G3, reason_rep samt reason_course.

Steg 2: Variablerna från steg 1 sattes in i en multivariat modell och de variablerna som fick ett p värde under 0.05 och behålls för nästa modell är sex, famrel, goout, absences. Därefter genomfördes ett likelihood ratio test mellan den större modellen med variablerna från steg 1 med den mindre modellen. Chivärdet vilket är differensen av vardera modells loglikelihood värde multiplicerat med -2 som blev 11.028. Differensen på frihetsgraderna 11 vilket gör att man kan räkna ut p-värdet som blev 0.441 vilket är över 0.05 vilket betyder att vi ej kan förkasta nollhypotesen alltså är de mindre modellen bättre passad.

Steg 3: När $\Delta\beta$ räknades ut kom det ut ett värde som var över 20% vilket indikerar på att de finns någon variabel som blivit borttagen som ger en signifikant skillnad. Därför kollades alla variabler som tagit bort från den större till den mindre genom att att använda samma likelihood ratio test. Det visade sig att inga variabel var kunde adderas så vi valde att gå vidare till steg 4 där man adderar de variablerna man valde bort i steg 1.

Steg 4: Efter att ha analyserat de borttagna variablerna från steg 1 så visade de sig att mom_at_home var signifikant för modellen och därför behålls den. Ett likelihood ratio test visade också att den modellen med mom_at_home var bättre. Detta är nu den preliminära huvudeffekts modellen och innehåller variablerna sex, famrel, goout, absences samt mom_at_home.

Steg 5: Variablerna sex och mom_at_home är binära alltså antar man att dessa är linjära. Men famrel, goout, absences måste kollas. I Figur 7 ser man dessa figurer och där ser man att de är linjära.



Figur 7: Logit-Transformationsplot

Steg 6: Samband studerades men ingen av de var signifikanta.

Steg 7: Dessa steg kommer genomföras längre ner i denna analys i samband med jämförlig av en annan modell.

Efter ha gjort en pruposeful selektion har man kvar en modell med endast de förklarande variablerna *sex*, *famrel*, *goout*, *absences*, *mom_at_home* denna modell heter ModellP2.

Stepwise:
I detta datasett ger även här both och backward metoderna exakt samma modell, denna modell kallar vi Backwards2 som innehåller variablerna *sex, famsize, famsup, paid, romantic, famrel, goout, health, absences, mom_at_home, dad_at_home*. Forward modellen innehåller väldigt många variabler med höga p-värden så denna fortsätter vi ej med. Även här selekteras möjliga samband bort när modellen framtogs.

VIF värdena kontrollerades för dessa modeller och de var bra så nu testar vi dessa två modeller i Tabell 10,11 och 12 för att se vilken som är bäst. I Tabell 10 utförs ett likelihood ratio test där modellerna testas mot en modell utan några förklarande variabler. Denna modell har alltså en frihetsgrad (df) samt ett loglikelihood värde på -140.80.

Modell	df	LogLik	χ^2	p-värde
ModellP2	6	-93.90	93.799	$< 2.2 \cdot e^{-16}$
Backwards2	12	-84.72	112.14	$< 2.2 \cdot e^{-16}$

Tabell 10: Likelihood ratio test

Det genomfördes även ett likelihood ratio test som syns i Tabell 11 som är mellan modellerna ModellP2 och Backwards2. Där kan man se att p-värdet är under 0.05 vilket betyder att nollhypotesen att den mindre modellen (ModellP2) är bäst kan förkastas.

Modell	Δdf	χ^2	p-värde
ModellP2/Backwards2	6	18.345	0.005426

Tabell 11: Likelihood ratio test mellan modellerna

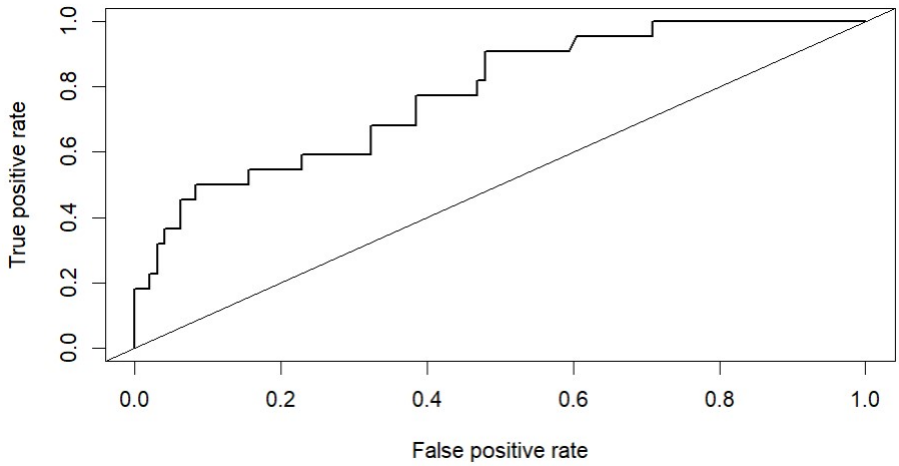
I Tabell 12 kollar man värdena av AIC och Pseudo R^2. Man ser där att Backwards2 har mindre AIC värde samt högre R^2 vilket betyder att Backwards2 är en bättre passning än ModellP2.

Modell	AIC	McFadden R^2	Cox&Snell R^2	Nagelkerke R^2
ModellP2	199.80	0.333	0.287	0.450
Backwards2	193.46	0.398	0.333	0.522

Tabell 12: AIC och R^2

Utifrån testresultaten i Tabell 10, 11 och 12 visar det att Backwards2 är bäst passande för datan.

ROC-kurvorna och AUC värdena jämfördes också där det visade sig att ModellP2 har ett AUC värde på 0.782 och Backwards2 ett AUC värde på 0.779. Värdena är väldigt nära varandra så trots att ModellP2 har ett lite högre AUC kommer Backwards2 fortsatt anses som den bättre modellen på grund av de andra testresultaten. ROC kurvan för Backwards2 presenteras i Figur 8. Datan som testade modellerna alltså för ROC-kurvan är den andra delen av datasetet som beskrivs i början av analysen, alltså 30 procent av den totala datan.



Figur 8: ROC-kurva för Backwards2

Den slutgiltiga modellen för responsvariabel 2 sammanställs i Tabell 13. Variabeln *dad_at_home* fick ett skattat värde -16.92 och ett standardfel på 1373.26 vilket tyder på att något fel har skett med modellen. Eftersom denna modell är gjort utifrån en backwards elimination gjort av programmet R så är det svårt att veta hur denna variabel blev en del av modellen. Detta är bra att ha i åtanke när diskussionen och slutsatsen tas. En modell med samma variabler som Backwards2 men utan *dad_at_home* testades men det gav inga större förändringar.

Koefficienter	Skattade värde	Standardfel	p-värde
Intercept	-3.35	1.28	0.0089
sex	-1.95	0.45	$1.71 \cdot e^{-5}$
famsize	-0.67	0.43	0.1179
famsup	-0.87	0.42	0.0384
paid	0.68	0.42	0.1092
romantic	-0.81	0.45	0.0680
famrel	-0.66	0.23	0.0040
goout	1.33	0.21	$5.65 \cdot e^{-10}$
health	0.26	0.16	0.0984
absences	0.057	0.023	0.0122
mom_at_home	1.74	0.56	0.0020
dad_at_home	-16.92	1373.26	0.99

Tabell 13: Slutgiltiga modellen för responsvariabel 2

Denna modell skrivs som Ekvation 19 där variabeln *sex* står för kvinna när $x_{sex} = 1$ och man när $x_{sex} = 0$, *famsize* står för familjestorlek där $x_{famsize} = 0$ står för mindre eller lika med tre och $x_{famsize} = 1$ står för större än tre i familjen. De numeriska variablerna är *famrel* där x_{famrel} kan ha värdena mellan 1-5 vilket representerar väldigt dålig till excellent kvalité av familjerelationer, *goout* där x_{goout} kan ha värdena 1-5 vilket representerar vara väldigt lite ute med vänner till väldigt mycket, *health* där x_{health} kan ha värdena 1-5 som representerar väldigt dålig till väldig bra hälsa och *absences* där $x_{absences}$ kan ha värdena mellan 0-93 vilket representerar antal frånvaro i skolan. Resterande variabler är binära variabler där $x = 1$ står för ja och $x = 0$ står för nej.

$$\begin{aligned} logit(\pi(x)) = -3.35 - 1.95x_{sex} - 0.67x_{famsize} - 0.87x_{famsup} + 0.68x_{paid} \\ -0.81x_{romantic} - 0.66x_{famrel} + 1.33x_{goout} + 0.26x_{health} + 0.057x_{absences} \\ +1.74x_{mom\_at\_home} - 16.92x_{dad\_at\_home} \end{aligned} \tag{19}$$

3.1.3 Responsvariabel 3

I denna modell framtagning kommer responsvariabeln vara satt att 1-4 är satt som låg alkoholkonsumtion och 5 är satt som hög alkoholkonsumtion. Fördelningen på variablerna är 367 stycken studenter som rapporterade med låg alkoholkonsumtion och 28 stycken med hög alkoholkonsumtion. Detta dataset delas in i två delar med ett dataset som innehåller 70 procent och en

med 30 procent, datasetdelen med 70 procent av den beskrivna datan kommer användas för modelleringen som nu kommer beskrivas hur den genomfördes.

Purposeful selection:

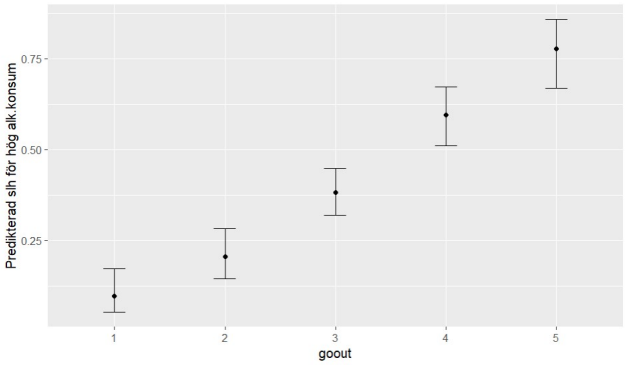
Steg 1: Efter första steget var de variablerna med ett p-värde under 0.25 sex, address, famsize, traveltime, studytime, failures, higher, famrel, freetime, goout, health och reason_rep.

Steg 2: Variablerna från steg 1 sattes in i en multivariat modell och variablerna i den modellen som har ett p-värde under 0.05 sätts in i en ny mindre modell. De variablerna som hade ett p-värde under 0.05 och hamnade i den mindre modellen blev endast sex samt goout. Ett likelihood ratio test mellan den större och mindre modellen visar att den större modellen passar bättre trots att den större modellen har väldigt höga p-värden. Vi går vidare till nästa steg.

Steg 3: I detta steg jämförs värdena mellan den större modellen och den mindre modellen. Inget av värdena hade ett $\Delta\beta > 20\%$ vilket betyder att ingen av de borttagna variablerna från den större till den mindre är viktiga. Därför fortsätter vi till steg 4.

Steg 4: Nu ska vi lägga tillbaka alla de variabler som togs bort i det första steget för att se om de är signifikanta. Det visade sig att mom_at_home kan läggas till i modellen. Detta betyder att den preliminära huvudeffektsmodellen innehåller variablerna sex, goout samt mom_at_home.

Steg 5: Både variablerna sex och mom_at_home är binära så de antas vara linjära. Variabeln goout behöver analyserat och detta ser man i Figur 9 där man ser att den är linjär.



Figur 9: Logit-Transformationsplot

Steg 6: Samband studerades men ingen av de var signifikanta.

Steg 7: Dessa steg kommer genomföras längre ner i denna analys i samband med jämförande av en annan modell.

Denna responsvariabel fick en modell ur purposeful selection som kallas ModellP3 och innehåller endast variablerna *sex, goout, mom_ at_ home.*

Stepwise:
Som tidigare analyser gav forward modellen väldigt många variabler med höga p-värden så den valdes bort direkt. Modellerna med both och backwards metoderna gav samma modell som vi kallar för Backwards3 och den innehåller variablerna *school, sex, age, famsize, paid, nursery, famrel, goout, mom_ health, mom_ at_ home, dad_ health.*

Efter modelleringen testades modellernas VIF värden och alla dessa var godkända. I Tabell 14,15 och 16 genomförs tester för att jämföra modellerna med varandra. I Tabell 14 utförs ett likelihood ratio test där varje modell testas mot en modell utan några förklarande variabler. Denna modell har en frihetsgrad (df) samt ett loglikelihood värde på -76.83. Tabell 14 visar att båda modeller får ett p-värde under 0.05 vilket betyder att man kan förkasta nollhypotesen vilket betyder att variablerna i modellerna är relevanta.

Modeller	df	LogLik	χ^2	p-värde
ModellP3	4	-54.27	45.122	$8.715 \cdot e^{-10}$
Backwards3	12	-44.30	65.048	$1.056 \cdot e^{-9}$

Tabell 14: Likelihood ratio test för ModelP3

Därefter genomfördes ett likelihood ratio test mellan modellerna för att se vilken av modellerna som är bättre. I Tabell 15 kan man se att p-värdet är mindre än 0.05 vilket betyder att den större modellen som är Backwards3 är den mest lämpade modellen.

Modell	Δdf	χ^2	p-värde
ModellP3/Backwards3	8	19.925	0.01062

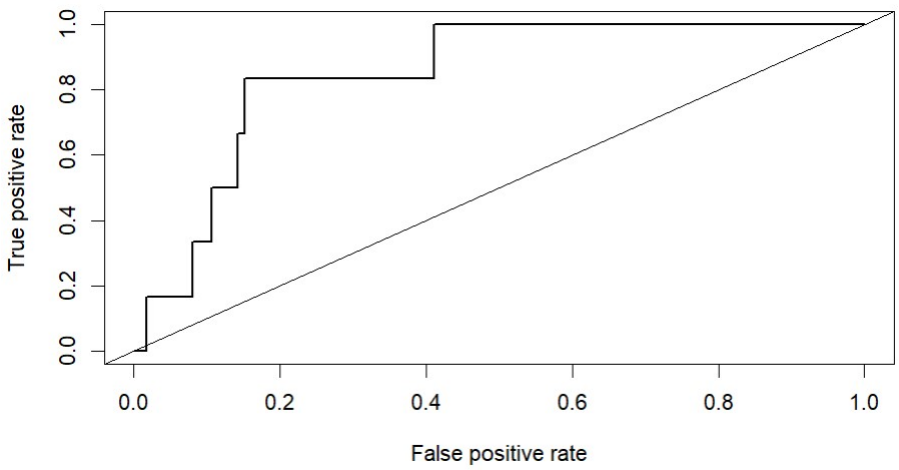
Tabell 15: Likelihood ratio test mellan modellerna

Tabell 16 visar modellernas AIC och Pseudo R^2 värden för vardera modell. Den visar att Backwards3 har lägre AIC samt högre R^2 vilket betyder att den är bäst lämpad.

Modeller	AIC	McFadden R^2	Cox&Snell R^2	Nagelkerke R^2
ModellP3	116.53	0.294	0.150	0.353
Backwards3	112.61	0.423	0.209	0.492

Tabell 16: AIC och R^2

ROC-kurvorna och AUC värdena jämfördes där ModellP3 fick ett AUC värde på 0.858 medans Backwards3 fick ett värde på 0.848. Värderna är alltså väldigt nära och kombinerat med testresultaten dras slutsatsen att Backwards3 är den mest lämpade modellen för denna data. I Figur 10 ser man ROC-kurvan för Backwards3. Datan som testade modellerna alltså för ROC-kurvan är den andra delen av datasetet som beskrivs i början av analysen, alltså 30 procent av den totala datan.



Figur 10: ROC-kurva för Backwards3

Den slutgiltiga modellen för responsvariabel 3 presenteras i Tabell 17.

Koefficienter	Skattade värde	Standardfel	p-värde
Intercept	-14.06	5.54	0.0111
school	2.71	1.11	0.0143
sex	-3.61	0.90	$6.44 \cdot e^{-5}$
age	0.49	0.28	0.0823
famsize	-1.56	0.63	0.0127
paid	1.21	0.61	0.0457
nursery	-1.36	0.72	0.0577
famrel	-0.66	0.33	0.0485
goout	1.49	0.34	$1.16 \cdot e^{-5}$
mom_health	-2.15	1.33	0.1049
mom_at_home	1.89	0.76	0.0125
dad_health	2.47	1.36	0.0699

Tabell 17: Slutgiltiga modellen för responsvariabel 3

Denna modell skrivs ut i Ekvation 20 där där variabeln *school* står för vilken skola eleven går i där $x_{school} = 0$ är skolan MS och $x_{school} = 1$ är skolan GP. Variabeln *sex* står för kvinna när $x_{sex} = 1$ och man när $x_{sex} = 0$, variabeln x_{age} står för vilket ålder eleven har i intervallet 15-22 och *famsize* står för familjens storlek där $x_{famsize} = 0$ står för en familj som är mindre eller lika med tre samt $x_{famsize} = 1$ står för en familj som är strikt större än tre. Variabeln *paid* står för extra betalda platser där $x_{paid} = 0$ står för inga och $x_{paid} = 1$ står för att eleven har extra betalda klasser. Variabeln *nursery* står för om eleven gick på förskola eller ej där $x_{nursery} = 1$ betyder att man gått på förskola och $x_{nursery} = 0$ står för att eleven ej gått på förskola. *Famrel* står för kvaliten på familjerelationerna där x_{famrel} kan ha värden mellan 1 till 5 vilket är från väldigt dålig till excellent. Modellen innehåller även *goout* där x_{goout} kan ha värdena 1-5 vilket representerar vara väldigt lite ute med vänner till väldigt mycket och *mom_health* står för hurvida mamman jobbar inom hälsovård ($x_{mom\_health} = 1$) eller ej ($x_{mom\_health} = 0$). Variabeln *mom_at_home* där $x_{mom\_at\_home} = 1$ står för att studentens mamma är en hemmamamma samt $x_{mom\_at\_home} = 0$ står för att studentens mamma inte är en hemmamamma. Sist har modellen *dad_health* där $x_{dad\_health} = 0$ står för att pappan ej arbetar inom hälsovård och $x_{dad\_health} = 1$ att pappan gör.

$$\begin{aligned} logit(\pi(x)) = -14.06 + 2.71x_{school} - 3.61x_{sex} + 0.49x_{age} - 1.56x_{famsize} + \\ 1.21x_{paid} - 1.36x_{nursery} - 0.66x_{famrel} + 1.49x_{goout} - 2.15x_{mom\_health} + \\ 1.89x_{mom\_at\_home} + 2.47x_{dad\_health} \end{aligned} \tag{20}$$

För att få ett tydligt resultat över vad dessa modeller gett kan man se i Tabell 18 en sammanfattning av de variabler som blev en del av samtliga modeller som blivit skapat utifrån Tabell 9,13 och 17.

Variabler	Backwards1	Backwards2	Backwards3
Intercept	-1.64	-3.35	-14.06
sex	-0.58	-1.95	-3.61
famsize	-0.53	-0.67	-1.56
famrel	-0.43	-0.66	-0.66
goout	0.88	1.33	1.49
absences	0.027	0.057	
dad_at_home	-1.63	-16.92	
paid		0.68	1.21
mom_at_home		1.74	1.89

Tabell 18: Gemensamma variabler mellan de slutgiltiga modeller

4 Diskussion

En slutgiltig modell har framtagits vardera från de tre olika responsvariablerna och ur Tabell 18 kan man se att de gemensamma variablerna som de tre modellerna har är *sex*, *famsize*, *famrel* och *goout*. Dessa variabler står för kön, familjens storlek, kvalitén på familjerelationerna och hur mycket studenter går ut med sina vänner. Variabeln *sex* står för vilket kön studenten har där $x_{sex} = 0$ betyder att de är en man och $x_{sex} = 1$ står för att de är en kvinna. I Tabell 18 kan man se alla skattade värden för de gemensamma variablerna modellerna har, där ser man att *sex* för samtliga modeller har ett negativt värde. Detta betyder att kvinnor har lägre sannolikhet att ha en hög alkoholkonsumtion än män. Detta stämmer överens med tidigare forskning som fått slutsatsen att män dricker mer än kvinnor. Dessa studier har dock visat att denna skillnad har minskat över tid (Folkhälsomyndigheten 2022). Variabeln *famsize* som står för storleken av studentens familj var en av de gemensamma variablerna och den var negativ. Detta betyder att om man har en större familj så är det en lägre sannolikhet att ha en hög alkoholkonsumtion. Detta skulle kunna bero på att med en större familj är man mindre ensam och om det skulle vara så att man har ett missbruk finns det fler som kan uppmärksamma det och fortare ge stöd. En annan av variablerna som de har gemensamt är *famrel* vilket står för kvalitén av familjerelationerna som studenten har. För alla modeller var denna variabel negativ vilket betyder att en bättre familjerelation leder till en lägre sannolikhet till att konsumera

mycket alkohol. Variabeln *goout* var även en variabel som alla modeller har gemensamt och denna står för hur mycket studenten går ut med sina vänner. Denna variabel var positiv för samtliga modeller vilket betyder att ju mer en student går ut med sina vänner desto högre sannolikhet är det att konsumera mycket alkohol. Eftersom denna studie har studenter som främst var mellan 15 till 18 år betyder det att för majoriteten som svarade på denna studie är konsumtion av alkohol olagligt. Studien utfördes som sagt i Portugal där åldersgränsen för alkoholkonsumtion är 18 år. Detta betyder alltså att gå ut med vänner främst inte handlar om att gå till barer på helgen. En studie gjord av IQ visade att minderåriga dricker alkohol på grund av sociala skäl för att det är kul samt för att våga men även för att man är rädd för att vara en tråkig person för att man inte vill dricka. Under tonåren är behoven starkare gällande att passa in, skaffa kompisar och slippa hamna i utanförskap. Samhällets normer och förväntningar ser alkohol som en naturlig del av livet eftersom det är runtomkring oss i samhället såsom i filmer, ute på restauranger, hemma och så vidare. Alltså är det inte konstigt att unga känner att det är förväntat av en att prova och se hur man beter sig när man är berusad. (IQ.se u. å.)

Mellan första och andra responsvariabeln hade de utöver de fyra variablerna som alla modeller har gemensamt två stycken gemensamma variabler. Dessa är *absences* samt *dad_at_home*. Variabeln *absences* står för hur mycket frånvaro en student har från skolan, denna hade ett skattat värde som är positivt vilket betyder att ju mer frånvaro desto högre sannolikhet är det att studenten har en hög alkoholkonsumtion. Den andra gemensamma är *dad_at_home* vilket är om ens pappa är en hemmapappa eller ej. Denna variabel är negativ vilket betyder att om man har en hemmapappa så är det mindre sannolikt att ha en hög alkoholkonsumtion. Detta samband är svår att förklara men olika anledningar kan diskuteras. I Portugal som denna data blivit hämtad från har en blandad hushållsstruktur. En artikel av Chara Scroope 2018 beskrevs Portugals hushållsfördelning som att i lantliga delar är pappan den som tar hand om familjens inkomster och mamman förväntas ta hand om hem och familj. Men i städerna är det mer jämt fördelat (Cultural Atlas 2018). I världen över är den sociala normen att om man har en förälder hemma är det kvinnan som ska göra hemmasysslorna. Alltså kan det vara så att om pappans roll i familjen är att vara hemma kan detta tyda på en hälsosam relation som inte tror på sterotypiska roller i hemmet och att antagligen mammas arbete ger hemmet en bra inkomst. Detta kan alltså vara en möjlig anledning till varför sannolikheten att ha en hög alkoholkonsumtion är lägre när man har en hemmapappa.

Denna analys kan även kopplas till den gemensamma variabeln som Backwards2 och Backwards3 har vilket är *mom_at_home* som är positiv vilket

betyder att om ens mamma är en hemmamamma så är sannolikheten högre att ha en hög alkoholkonsumtion för studenten. Detta skulle kunna vara eftersom de sociala normerna tycker att mamman ska vara hemma kan det finnas kvinnor som är hemmamammor när de hellre skulle vilja vara i arbetslivet. Studier har visat att hemmamammor är mer olyckliga, irriterade och deprimerade än jobbande mammor (Nomaguchi och Milkie 2003). Utifrån denna studie kan de betyda att en student med en hemmamamma lever i en mer olycklig familj vilket kan leda till att studenten vänder sig till alkohol.

Ytterligare variabel som andra och tredje responsvariabeln har gemensamt är *paid* som är positiv, detta betyder alltså att om eleven har extra betalda klasser är sannolikheten högre att ha en hög alkoholkonsumtion. Detta skulle kunna bero på att eleven får extra mycket press lagt på sig och har svårt att hitta tid att koppla av. Det är även ej specificerat i databeskrivning om dessa extra betalda klasser är mer avancerade klasser eller för elever som behöver mer stöd. Alltså är det svårt att lägga en tydlig hypotes vad detta kan betyda.

Storleken på de olika koefficienterna påverkar hur stor påverkan de förklarande variablerna har på responsvariabeln. Variabeln *sex* har ett genomsnitt koefficient värde mellan de olika modellerna på -2.05 och en annan binär variabel *mom_at_home* hade ett genomsnitt koefficient värde på 1.82 vilket betyder att dessa variabler påverkar ungefär lika mycket men åt olika håll. Ena ökar sannolikheten och den andra minskar sannolikheten för hög alkoholkonsumtion. De binära variablerna *famsize* och *paid* har koefficient värde på -0.92 respektive 0.945 vilket betyder att de också påverkar lika mycket fast åt olika håll.

Variabeln *dad_at_home* hade intressanta koefficientvärden då värdet från Backwards2 är problematisk på grund av hur stor den är, värdet på den som man ser i bland annat Tabell 18 är -16.92 medan i Backwards1 är koefficientvärdet -1.63. Antar man att Backwards2 ger ett felaktigt svar för *dad_at_home* så utgår man på Backwards1 värde och detta gör att den påverkar lite mindre än *sex* och *mom_at_home*. En anledning till att man får ett så stort värde på *dad_at_home* kan bero på dess fördelning vilket kan synas i Figur 11 i appendix. Den visar att det är väldigt få som faktiskt har en hemmapappa vilket kan vara en förklaring till varför responsvariabel 2 i Tabell 13 visar sig ha ett skattat värde på -16.92 och det stora standardfelet på 1373.26. Anledningen till att responsvariabel 1 i Tabell 9 har ett skattat värde på *dad_at_home* på -1.63 med ett standardfel på 0.74 kan vara att många som har en pappa hemma kan vara de som har en alkoholkonsumtion på grad tre. Responsvariabel är uppdelad så att på en skala 1-5 anses 1-2 som låg alkoholkonsumtion och 3-5 en högalkohol konsumtion medans responsvariabel 2 är uppdelad så att 1-3 ses som låg alkoholkonsumtion och

4-5 som en hög alkoholkonsumtion. Skillnaden på dessa responsvariabler är alltså huruvida man definierar grad 3 som en hög eller låg alkoholkonsumtion. Undersöker man närmare på fördelningen mellan *Walc* och *dad_at_home* i Tabell 19 och 20 i appendix att fördelningen med responsvariabel 1 är mer likafördelat än responsvariabel 2. Tabell 20 visar att det ej finns någon observation där man har en hög alkoholkonsumtion samt en pappa hemma. Alltså är alla möjliga händelser ej representerade till skillnad från Tabell 19 vilket kan vara anledningen till att skattningen för *dad_at_home* har så hög skattat värde och så högt standardfel.

Eftersom variabeln *dad_at_home* analyserades och vi har tidigare jämfört denna variabel med *mom_at_home* är det rimligt att granska fördelning med. I Figur 12 ser man att jobb mellan mammorna är mycket mer utspritt. I Tabell 21 och 22 i appendix kan man även se hur de finns fler observationer där studenterna har en mamma hemma vilket gör att skattningen och standardfelet är mer pålitligt än för *dad_at_home*.

Ytterligare variabel är *goout* vilket har ett genomsnittligt koefficientvärde på 1.23 men detta är en numerisk variabel så detta värde kan multiplicerat med 1,2,3,4 eller 5. Alltså är den minimala påverkan *goout* har på responsvariabeln Walc $1.23 \cdot 1 = 1.23$ och den maximala påverkan är $1.23 \cdot 5 = 6.15$. Var dock medveten om att detta värde är ett medelvärde för att få en överblick av variabeln påverkas. En annan numerisk variabel som kan multipliceras med 1,2,3,4 eller 5 vilket är *famrel* och denna har mellan de olika modellerna ett medelvärdes koefficientvärde på -0.58 vilket i första anblick är lågt. Men detta är de lägsta påverkan variabeln kan ha på responsvariabeln, det maximala värden är $-0.58 \cdot 5 = -2.9$. Slutligen finns variabeln *absences* som har ett medelvärdes koefficientvärde på 0.042, denna variabel är numerisk mellan 0-93 vilket betyder att minsta påverkan variabeln kan ha är 0 och det maximala är $0.042 \cdot 93 = 3.906$. Det är alltså svårt att genom att bara titta på koefficientvärdena avgöra variablernas påverkan, men de ger en överblick.

En viktig del av denna analys är att datan är från Portugal som har en annan kultur än Sverige. Enligt en rapport på 15-16 åringar från European Monitoring Centre for Drugs and Drug Addiction (ESPAD) så hade i Portugal 77 procent druckit alkohol i sitt liv och 43 procent hade druckit alkohol de 30 senaste dagarna. I samma rapport med 15-16 åriga studenter i Sverige hade 58 procent druckit alkohol i sitt liv och 25 procent hade druckit de 30 senaste dagarna (Mokinaro m. fl. 2020). Detta visar att det är vanligare i Portugal att ungdomar dricker i tidig ålder och att det ligger mer i deras kultur än i Sverige.

När man gör analyser finns det alltid felkällor, datan för denna rapport är baserad på en enkät som studenter har svarat på. Det finns alltså en risk att svaren som skickas in inte är ärliga svar och detta kan bero på olika saker

såsom man ej vill svara eller kanske skäms över sina svar. Alltså är det viktigt att vara medveten om detta när man presenterar slutsatsen då den kan ha felaktig data. Om denna statistiska analys skulle göras om så skulle det vara en idé att använda sig av bootstrap som är en statistisk metod som generar nya datapunkter utifrån de man har.

Om man skulle ha möjlighet att utveckla sin rapport så skulle en bra ide vara att prova olika ‘set.seed()‘ när man slumpar fram datamängden för modellering och testen vilket var uppdelning med 70 procent respektive 30 procent. Det hade varit intressant att se om man skulle få samma slutsats om man använde olika ‘set.seed()‘ och i så fall se vad skillnaden är. Detta skulle även vara relevant eftersom det var väldigt många variabler som valdes bort så det skulle vara intressant att se vilka variabler som valdes bort av slumpen och vilka variabler som har ett starkt samband till responsvariabler oberoende vilket slumpmässigt dataset som väljs. Detta gjordes med en hastig undersökning för att få en överblick över skillnaderna som analyserna kan ge. Det visade att de mest signifikanta variablerna som representerade kön (*sex*) och hur mycket studenten går ut (*goout*) var väldigt signifikanta i alla dataset. De variabler som visade sig vara signifikanta i analyserna med responsvariabel 1 och 2 i nästan varje ´set.seed()´ var samtliga gemensamma variabler som resultatet gav. Variabeln som ej diskuterades för dessa resultat som kom upp i andra analyser var *health* vilket står för studentens nuvarande hälsotillstånd. Denna lilla undersökning gjordes med backwards metoden eftersom de var den metoden som gav den bästa modellen för samtliga responsvariabler. En mer utförlig analys skulle undersöka även andra metoder som purposeful selektion.

I denna rapport användes två olika metoder, stepwise samt purposeful selection men alla slutgiltiga modeller kom utifrån backwards eliminering. Detta kan bero på att backwards eliminering är en väldigt känd metod som anpassar modellen till datan. Den tar fram den modell som får lägst AIC värde vilket betyder att under samma tid som modellen anpassas testas den också. En stepwise eliminering är även en mer strikt metod än purposeful selection. Purposeful selectionen var mycket mer omständig med många steg och många variabler att ha kontroll på så det ökar risken för misstag i analysen. För responsvariabel 2 och 3 hade modellerna från purposeful selection dock en bättre prediktionsförmåga, det var dock en väldigt liten skillnad. Men eftersom huvudsyftet var att hitta möjliga faktorer som leder till hög alkoholkonsumtion så var det mer viktigt att modellen passar datan viktigare.

Idéer för framtida projekt skulle vara att främst göra detta med svensk data för att se om det skulle få samma förklarande variabler som med detta dataset. Det skulle även vara intressant att undersöka med andra variabler såsom psykologiska faktorer, föräldrars alkoholvanor, gener och så vidare.

Detta dataset samlades in 2008 och det skulle vara intressant att göra analysen på nytt eftersom alkoholkulturen i samhället ändras under tid. Det skulle även vara intressant och mer relevant att ha en responsvariabel som representerar vardaglig alkoholkonsumtiom. Detta fanns i detta dataset men hade tyvärr för få datapunkter med hög alkoholkonsumtion för att kunna göra en bra analys. Ett alternativ att lösa detta skulle kunna vara under eller översampling, där undersampling är när man slumpmässigt tar bort punkter från den överrepresenterade klassen och översampling är när man ökar punkter i den underrepresenterade klassen genom att exempelvis duplicera befintliga punkter.

Slutsatsen som kan dras utifrån denna rapport är att kön, hur mycket ungdomar går ut samt familjens storlek och relationer har ett samband med hur mycket alkohol som studenten konsumerar. Det är viktigt att informera tonåringar i tid om hur alkohol påverkar kroppen och hur missbruk kan vara genetiskt, vilket gör att beroende av alkohol är lättare för vissa människor att få på grund av genetik.

5 Källförteckning

Agresti, Alan (2002). *Categorical Data Analysis.* Vol. 2. Wiley-Interscience, s. 24, 88, 122, 165–166, 178, 186, 213, 216.

— (2013). *Categorical Data Analysis.* Vol. 3. Wiley-Interscience, s. 19, 212.

Cortez, P. och A. Silva (2008). *Student Alcohol Consumption.* URL: `https://www.kaggle.com/datasets/uciml/student-alcohol-consumption?resource=download&select=student-merge.R`.

Craney, Trevor A och James G Surles (2002). "Model-dependent variance inflation factor cutoff values". I: *Quality engineering* 14.3, s. 391–403.

Cultural Atlas (2018). *Portuguese Culture - Family.*

Folkhälsomyndigheten (2022). "Alkoholens skadeverkningar". I.

Foroud, Tatiana, Howard J Edenberg och John C Crabbe (2010). "Genetic research: Who is at risk for alcoholism?" I: *Alcohol Research & Health* 33.1-2, s. 64.

Guttormsson, Ulf (2020). "Skolelevers drogvanor 2020". I: *CAN: Stockholm, Sweden.*

Hardy, Melissa A (1993). *Regression with dummy variables.* Vol. 93. Sage, s. v.

Hastie, T. J. och D. Pregibon (1992). *Generalized linear models.*

Hosmer Jr, David W, Stanley Lemeshow och Rodney X Sturdivant (2013). *Applied logistic regression.* Vol. 398. John Wiley & Sons, s. 8, 9, 10, 35, 162.

IQ (2021). "Svenskarnas inställning till vardagsdrickande". I: *Vardagsindex 2021*, s. 4.

IQ.se (u. å.). *Varför dricker tonåringar?* `https://www.iq.se/tonarsparloren/varfor-dricker-tonaringar/`. Åtkomstdatum: 7/5/2023.

Mokinaro, Sabrina m. fl. (2020). "ESPAD Report 2019: Results from European school survey project on alcohol and other drugs". I.

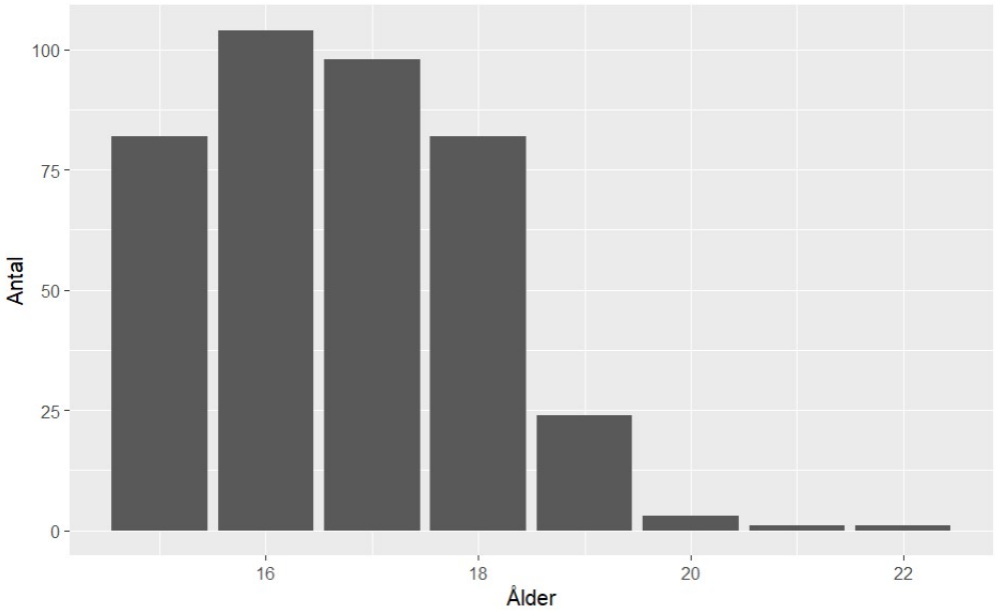
Muschelli III, John (2020). "ROC and AUC with a binary predictor: a potentially misleading metric". I: *Journal of classification* 37.3, s. 696–708.

Nomaguchi, Kei M och Melissa A Milkie (2003). "Costs and rewards of children: The effects of becoming a parent on adults' lives". I: *Journal of marriage and family* 65.2, s. 356–374.

Ruengvirayudh, Pornchanok och Gordon P Brooks (2016). "Comparing stepwise regression models to the best-subsets models, or, the art of stepwise". I: *General linear model journal.*

Walker, David A och Thomas J Smith (2016). "Nine pseudo R2 indices for binary logistic regression models". I: *Journal of Modern Applied Statistical Methods* 15.1, s. 848–854.

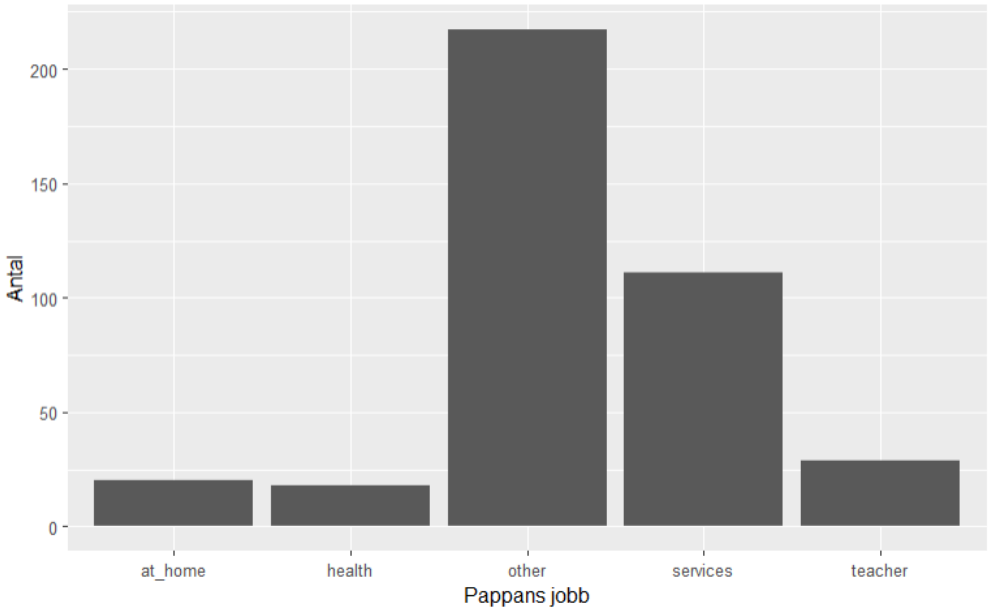
6 Appendix A



Figur 1: Åldersfördelning

Variabel	Nivåer
school	0 - MS, 1 - GP
sex	0-Man, 1-Kvinna
address	0 - Lantlig, 1 - Urban
famsize	0 - LE3, 1 - GT3
Pstatus	0 - Isär, 1 - Tillsammans
schoolsup	0 - Nej, 1 - Ja
famsup	0 - Nej, 1 - Ja
paid	0 -Nej, 1 - Ja
activities	0 - Nej, 1 - Ja
nursery	0 - Nej, 1 - Ja
higher	0 - Nej, 1 - Ja
internet	0 - Nej, 1 - Ja
romantic	0 - Nej, 1 - Ja
mom_teacher	0 - Nej, 1 - Ja
mom_health	0 - Nej, 1 - Ja
mom_services	0 - Nej, 1 - Ja
mom_at_home	0 - Nej, 1 - Ja
dad_teacher	0 - Nej, 1 - Ja
dad_health	0 - Nej, 1 - Ja
dad_services	0 - Nej, 1 - Ja
dad_at_home	0 - Nej, 1 - Ja
reason_home	0 - Nej, 1 - Ja
reason_rep	0 - Nej, 1 - Ja
reason_course	0 - Nej, 1 - Ja
guardian_mom	0 - Nej, 1 - Ja
guardian_dad	0 - Nej, 1 - Ja

Tabell 4: Transformerade variabler



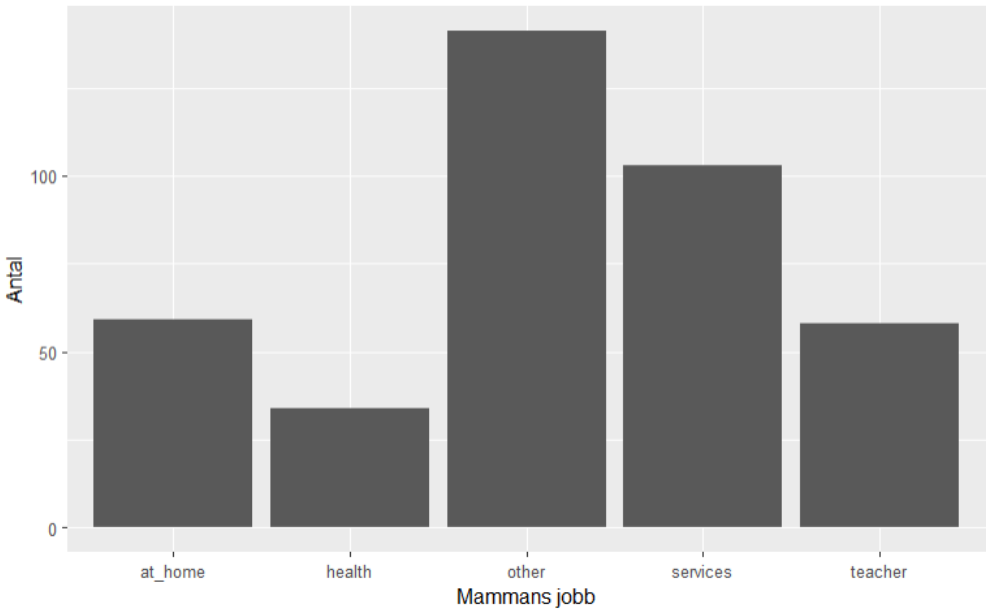
Figur 11: Fördelning pappans jobb

	dad_at_home	
Walc	0	1
0	219	17
1	156	3

Tabell 19: Responsvariabel 1

	dad_at_home	
Walc	0	1
0	296	20
1	79	0

Tabell 20: Responsvariabel 2



Figur 12: Fördelning mammans jobb

	mom_at_home	
Walc	0	1
0	270	46
1	66	13

Tabell 21: Responsvariabel 2

	mom_at_home	
Walc	0	1
0	314	53
1	22	6

Tabell 22: Responsvariabel 3