Stockholms universitet

Framgångsfaktorer för slutbetyget i gymnasiet

Henrik Norelius

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2017:16
Matematisk statistik
Juni 2017

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2017:16**
http://www.math.su.se/matstat

Framgångsfaktorer för slutbetyget i gymnasiet

Henrik Norelius*

Juni 2017

Sammanfattning

Syftet med denna uppsats är att på ett enkelt, översiktligt och konstruktivt sätt förklara en elevs slutbetyg från gymnasieskolan. Några av variablerna är kända innan elevens skolgång, exempelvis föräldrarnas utbildningsnivå, emedan andra variabler är en del av individens skolgång, exempelvis matematikbetyget i gymnasiet.
Vi kommer att jämföra två modeller varav en, den linjära regressionsmodellen, kommer anses olämplig och väljas bort till förmån för en logistisk regressionsmodell. Som kan tyckas väntat så finner vi att slutbetyget i årskurs 9 spelar en stor roll för hur bra eleven presterar i gymnasiet, men också att de olika gymnasieprogrammen har stor inverkan på förväntat studieresultat.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: henrik.norelius@gmail.com. Handledare: Mathias Lindholm.

Innehåll

Sammanfattning

Syftet med denna uppsats är att på ett enkelt, översiktligt och konstruktivt sätt förklara en elevs slutbetyg från gymnasieskolan. Några av variablerna är kända innan elevens skolgång, exempelvis föräldrarnas utbildningsnivå, emedan andra variabler är en del av individens skolgång, exempelvis matematikbetyget i gymnasiet.
Vi kommer att jämföra två modeller varav en, den linjära regressionsmodellen, kommer anses olämplig och väljas bort till förmån för en logistisk regressionsmodell. Som kan tyckas väntat så finner vi att slutbetyget i årskurs 9 spelar en stor roll för hur bra eleven presterar i gymnasiet, men också att de olika gymnasieprogrammen har stor inverkan på förväntat studieresultat.

1 Introduktion

Under vårterminen 2014 tog 48 322 personer en gymnasieexamen från ett högskoleförberedande gymnasieprogram. De högskoleförberedande programmen är konstruerade så att de förbereder eleverna för att studera vidare på högskolan och är således mer teoretiskt inriktade än de övriga programmen. I vår analys ingår alltså elever som erhållit en gymnasieexamen från Samhällskunskap-, Ekonomi-, Naturvetenskap-, Teknik-, Humanist- eller Estetprogrammen. Trots olika programinriktningar kan sägas att det är en hyfsat homogen population.
Datamaterialet har vi erhållit från Skolverket och det är fullständigt anonymiserat.

I kapitel 2 kommer vi utförligt beskriva alla våra variabler mha. tabeller och figurer, och vi kommer även utföra viss databehandling inför vår analys. Kapitel 3 innehåller en stor del av den teori som sedan används i analysen. Inga bevis finns i kapitlet, utan istället ges förklaringar och vissa hänvisningar. I kapitel 4 gör vi vår statistiska analys; två modeller kommer härledas och presenteras, varav en kommer att anses som lämplig för vårt ändamål. Målsättningen i detta kapitel är att på ett överskådligt och tydligt sätt förklara och motivera vad vi gör. I kapitel 5 kommer främst den slutgiltiga modellen diskuteras och personliga tolkningar kommer att göras till varför resultatet kan tänkas se ut som det gör. Avslutningsvis finns det med ett avsnitt med förslag till förbättringar.

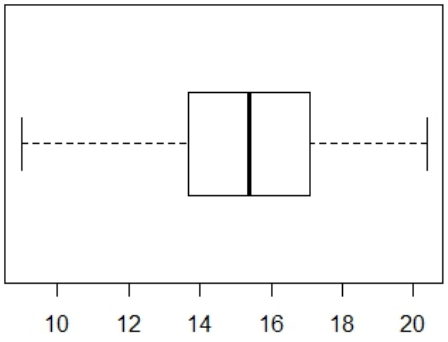
2 Deskriptiv Analys & Databehandling

2.1 Responsvariabel, Slutbetyg

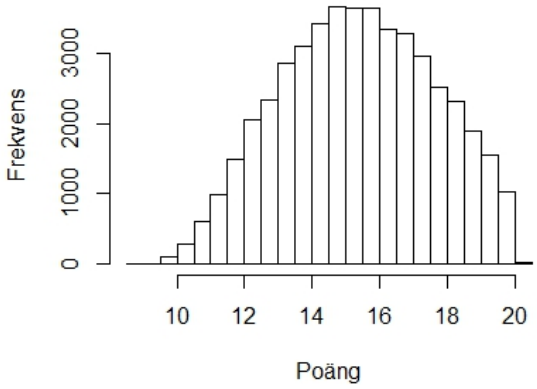
Variabel	Typ	Beskrivning
slutbetyg	Numerisk	Studentens slutbetyg i gymnasiet, värde mellan 9 - 20 poäng.

Slutbetyg är det genomsnittsbetyg som studenten får när denne har tagit en gymnasieexamen. I varje gymnasiekurs får eleven ett betyg, och slutbetyget utgör det viktade snittbetyget av samtliga enskilda betyg. Betygen som kan erhållas i varje kurs är A (20 p), B (17.5 p), C (15 p), D (12.5 p), E (10 p) och F (0 p). De fem första betygen erhåller eleven om denne klarar kursen i fråga, och betyget F är ett icke-godkänt betyg. Varje betygssteg har ett numeriskt värde som kallas "betygspoäng" vilket står inom parantes efter respektive betyg.
Vidare har varje kurs en omfattning som utgörs av ungefärligt antal timmar som kursen har schemalagd tid, dessa refererar vi till som "omfattningspoäng". Betygspoängen multipliceras med respektive omfattningspoäng för varje kurs, och sedan divideras summan med totalt antal omfattningspoäng, dvs. vi får formeln $\sum \frac{O_i \cdot B_i}{K}$, där B och O står för betygspoäng respektive omfattningspoäng för specifik kurs och K står för totala antalet omfattningspoäng som eleven läst. Resultatet blir ett viktat genomsnittspoäng som alltså utgör studentens slutbetyg. För att få en gymnasieexamen och ett slutbetyg skall en elev ha läst minst 2500 omfattningspoäng, varav minst 2250 måste utgöras av godkända betyg. Studenten har möjlighet att läsa fler kurser om denne så önskar, förutsatt att respektive skola kan tillgodose detta.

I vår analys kommer vi att justera slutbetyget genom att subtrahera matematikbetygets inverkan på variabeln, detta gör vi för att minska korrelationen mellan slutbetyg och variablerna som bygger på matematikbetyget i gymnasiet. Se figur 1 och 2 för boxplot respektive histogram av slutbetyget nedan.



Figur 1: Boxplot av slutbetyg



Figur 2: Histogram av slutbetyg

2.2 Förklaringsvariabler

I detta kapitel kommer vi presentera och analysera våra förklaringsvariabler, samt utföra viss databehandling på dem. I diagrammet nedan finns de översiktligt beskrivna.

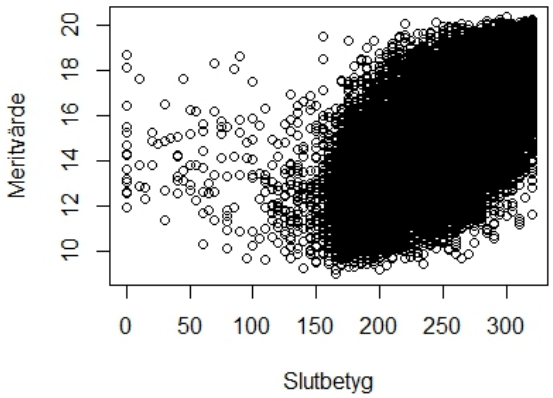
Variabel	Typ	Beskrivning
Meritvärde	Numerisk	Studentens meritvärde från grundskolan
Kön	Kategorisk	Elevens kön
Utländsk Bakgrund	Kategorisk	Anger om eleven eller elevens föräldrar är födda utomlands eller ej
Program	Kategorisk	Gymnasieprogrammet eleven tog examen inom
Föräldrarnas utbildningsnivå	Kategorisk	Den högst utbildade förälderns utbildning
Friskola	Kategorisk	Anger om skolan är en kommunal- eller friskola
Matgr	Numerisk	Elevens betygspoäng i grundskolematematik
Mat1	Numerisk	Elevens betygspoäng i gymnasiets första matematikkurs
Mat123	Numerisk	Genomsnittligt betygspoäng i (upp till) de tre första matematikkurserna, gymnasiet
Matsnitt	Numerisk	Genomsnittligt betygspoäng i samtliga matematikkurser (som eleven läst) i gymnasiet
Studieomfattning	Kategorisk	Anger om eleven har läst utökat antal kurser
Storstad	Kategorisk	Anger om skolan ligger i ett storstadområde

2.2.1 Meritvärde

Variabel	Typ	Beskrivning
meritvarde	Numerisk	Studentens genomsnittliga betyg i grundskolan, anges från 7,5 - 20.

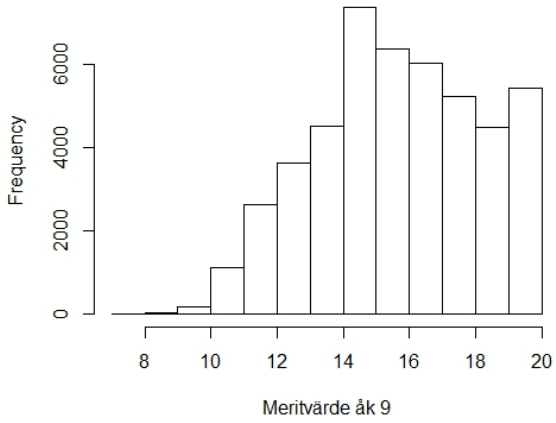
I årskurs nio får elever ett slutbetyg i varje ämne de har läst. Årskull 2011 gick ut med det gamla betygssystemet, dvs. MVG (20 p), VG (15 p), G (10 p) och IG (0 p). Eleven kan också få ett streck ”-” (0 p) istället för ett betyg, vilket betyder att underlag saknas. De två sistnämna utgör ej godkända betyg. Eleven får räkna in sin 16 bästa betyg, dessa summeras ihop och eleven får då ett värde mellan 0-320 vilket vi kallar meritvärde (den ”merit” som eleven söker in med, och konkurrerar sinsemellan, till gymnasieskolorna).
I denna studie vill vi se hur viktiga studentens meriter från grundskolan är för dennes möjlighet att få ett gott betyg i gymnasiet. Dock skall nämnas att för behörighet till ett högskoleförberedande gymnasieprogram (dvs. den typ av program studenterna vår population läser) krävs ett godkänt betyg i minst 12 ämnen, dvs. minst 120 poäng. Gissningsvis har de elever med ett lägre meritvärde sökt in med andra meriter, kompletterat upp betyg vid senare tillfälle eller läst ett år extra. Eftersom de har sökt in via sitt meritvärde kan detta värde inte jämföras med resten

av populationens värde. Vi kan anta att siffran 0 i meritvärde kan likställas med att inte få ett slutbetyg alls eftersom man då ej är behörig att söka gymnasiet. Vi kommer alltså att utesluta alla elever som har under 120 poäng i meritvärde.
Som kan ses i figur 3 så är det en hel del av populationen som inte uppfyller dessa krav. Dessutom kan vi se att det antagande om linjaritet mellan respons- (slutbetyg) och förklaringsvariabler inte verkar gälla så meritvärdet är runt 150 eller lägre.

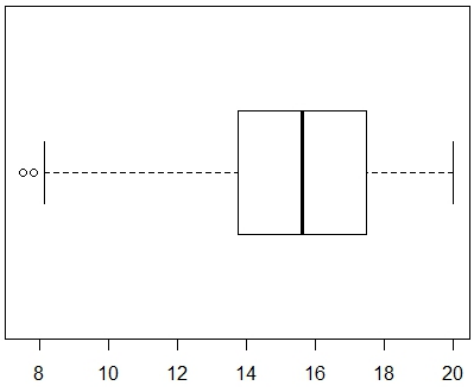


Figur 3: Plot av meritvärde och slutbetyg

I syfte att minska korrelationen mellan variabeln Matematikbetyg i grundskolan och Meritvärde skulle man kunna separera matematikbetyget i grundskolan från meritvärdet och på så vis få de 15 bästa betygens medelvärde. Dock vet vi inte om studenten i fråga har räknat med sitt matematikbetyg i meritvärdet, dvs. det kan vara så att eleven inte har matematikbetyget som ett av sina 16 bästa betyg. Detta skulle eventuellt kunna gå att räkna ut eller jaga data på, men vi väljer att ta det säkra före det osäkra och inte justera variabeln.
För att lättare tolka koefficienterna från meritvärdet kommer vi att formatera variabeln genom att dividera meritvärdet med 16 och då få en siffra mellan 7,5 (motsvarande 120) och 20 (motsvarande 320). Alltså samma format som variabeln *slutbetyg* anges på. Fördelningen av meritvärdet efter formatering och rensning av saknade värden (693 st) kan ses som boxplot i figur 5 samt 4 nedan.



Figur 4: Histogram av Meritvärde



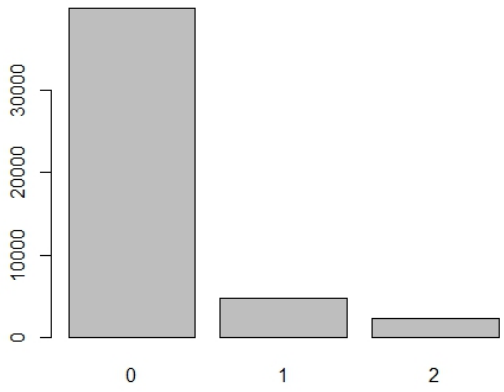
Figur 5: Boxplot av Meritvärde

2.2.2 Utländsk bakgrund

Variabel	Typ	Beskrivning
utlbakg	Kategorisk	0=“Inget utländskt påbrå”, 1=“Båda föräldrar födda utomlands” eller 2=“Eleven född utomlands”

Eftersom kulturell bakgrund, erfarenhet av olika skolsystem, förväntningar, studieteknik och eventuell studievana kan variera från Sverige och andra länder kan detta tänkas vara en intressant variabel att titta på. En majoritet av eleverna har

ingen utländsk bakgrund. Variabeln är behäftad med ett bortfall om 15 st fall. De utfallen som finns registrerade är illustrerade i figur 6

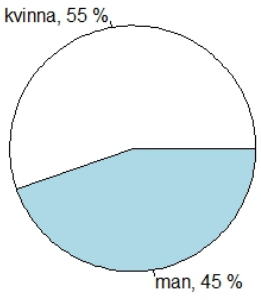


Figur 6: Utländsk bakgrund

2.2.3 Kön

Variabel	Typ	Beskrivning
kon	Kategorisk	Anger elevens kön; 0=“kvinna”, 1=“man”

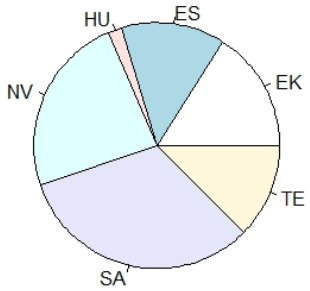
Det debatteras flitigt om kön kan tänkas vara kopplat till förväntat studieresultat och då vi data tillhandahållen har vi valt att undersöka om så är fallet.
Fördelning kan ses i pajdiagrammet nedan, figur 7.



Figur 7: Könsfördelning

2.2.4 Program

De olika programmen läser lite mer än 1000 undervisningstimmar gemensamma kurser, de s.k. gymnasiegemensamma ämnen (matematik 1) därefter skiljer sig de kurser en elev läser mer eller mindre åt efter vilket program eleven är inskriven på. Vårt datametrial innehåller elever som tagit examen från ett högskoleförberedande program, dvs. Ekonomi-, Humanist-, Naturkunskap-, Teknik-, Samhällskunskaps- eller Estestiskt program. Man kan alltså påstå att detta är en relativt homogen grupp att utföra analys på, däremot kan de ämnen som eleven läser skilja sig rätt mycket mellan vissa program. Se figur 8 för fördelning.



Figur 8: Fördelning av Programmen

Kategori	Beskrivning	Frekvens
SA	Samhällskunskapsprogrammet	15 209
EK	Ekonomiprogrammet	7 496
ES	Estetprogrammet	6 452
HU	Humanistprogrammet	801
NV	Naturkunskapsprogrammet	11 209
TE	Teknikrogrammet	5 861

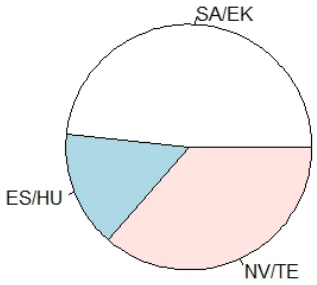
Tabell 1: Fördelning av program innan sammanslagning

När vi väl startar vår analys kommer vi dock inte att få signifikans för vissa program. Vi kommer då att angripa detta genom att göra en ihopslagning av programmen. De mest homogena programmen så tillvida att dels innehållet är det mest kongruenta/likvärdiga samt att de ”typiskt” sett leder till liknande behörigheter när man skall söka vidare till högre utbildning är att slå ihop programmen enligt nedanstående tabell.

Kategori	Beskrivning	Frekvens
SA/EK	Samhällskunskap- & Ekonomiprogrammet	22 705
ES/HU	Estet- & Humanistprogrammet	7 253
NV/TE	Natur- & Teknikrogrammet	17 070

Tabell 2: Fördelning av program efter sammanslagning

Fördelningen kan ses i pajdiagrammet 9 nedan. Med denna sammanslagning blir modellen enklare att tolka, som vi sätter vikt vid i denna analysen. Dessutom är blir grupperna mer likvärdiga storleksmässigt, vilket å ena sidan inte är superviktigt då vi har tillräckligt stora populationer i varje grupp, men fördelningen blir lite jämnare.

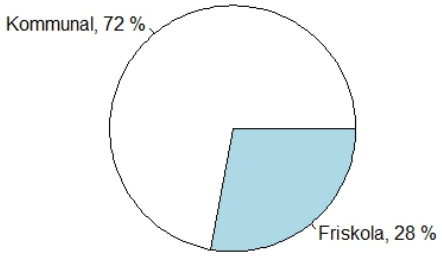


Figur 9: Fördelning av Programmen efter sammanslagning

2.2.5 Friskola

Namn	Typ	Beskrivning
$friskola$	Kategorisk	0= Kommunal 1=Friskola

Variabeln $friskola$ indikerar om en skola drivs av en kommun eller av ett annat organ. Om skolan inte drivs av en kommun drivs den av en annan juridisk person, i de flesta fall är detta ett Aktiebolag, men även andra former av juridiskapersoner förekommer. I denna studie skiljer vi bara på om den juridiska personen som driver skolan är en kommun eller inte. Vi kallar detta att skolan är ”kommunal” eller en ”friskola”. Generellt sett kan sägas att friskolor har lite mer frihet i sitt sätt att utforma undervisningen, men den stora skillnaden är att de har ett vinstintresse. En klar majoritet av eleverna i datamaterialet har studerat vid en kommunal skola, detta illustreras i figur 10.

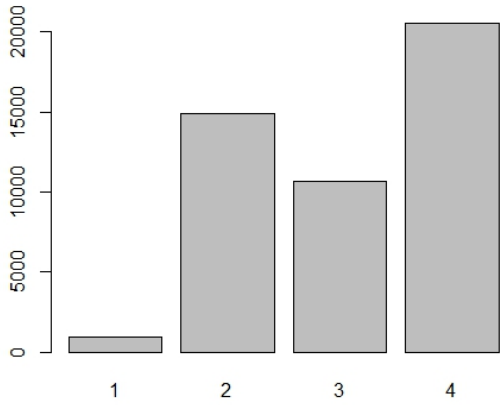


Figur 10: Fördelning kommunal och friskola

2.2.6 Föräldrarnas utbildningsnivå

Namn	Typ	Beskrivning
foruniva	Kategorisk	Den högst utbildade förälderns utbildning 0=“Förgymnasial”, 1=“Gymnasial”, 2=“Eftergymnasial <3år” eller 3=“Eftergymnasial ≥3år”

Det är rimligt att anta att en elevs inställning till sin skolgång påverkas av föräldrarnas inställning till skolan; om en elevs förälder tycker att utbildning är viktig är det troligtvis mer sannolikt att eleven också kommer värdera utbildning och ta sin skolgång på ett större allvar. Detta kan vara en viktig faktor för elevens studieresultat eftersom föräldrarna kan sporra och inspirera eleven att försöka få ett bra studieresultat. Vidare kan studievana i hemmet vara av vikt; om en elevs föräldrar själv har gått en högre utbildning kan denne lära eleven god studievana. Fördelningen kan ses i figur 11. Det är värt att notera att vi ser ett s.k. urvalsbias då fördelningen för utbildning i Sverige i stort inte ser ut på detta sätt; enligt Statistiska Centralbyrån [2] så har 25 % av befolkningen en eftergymnasial utbildning på 3 år eller längre och närmare 50 % har som högst en utbildning på gymnasial nivå. Detta är dock för enskilda individer och vårt mätvärde är den högst utbildade föräldern i ett hushåll, så vi har inte direkt samma frågeställning, men noterar att det ser ut som att de elever som ingår i vår studie har mer välutbildade föräldrar än den genomsnittliga utbildningsnivån i Sverige. I denna variabel finns också ett bortfall om 511 fall där vi alltså inte fått svar på vad föräldrarna har för utbildning.

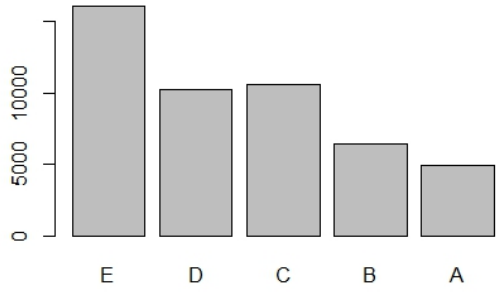


Figur 11: Föräldrarnas utbildningsnivå

2.2.7 Matematikbetyg kurs 1, gymnasiet

Namn	Typ	Beskrivning
*mat*1	Numerisk	Anger betygspoängen i den första matematikkursen i gymnasiet (respektive betygssteg inom parantes). 20 (=A), 17.5 (=B), 15 (=C), 12.5 (=D), 10 (=E).

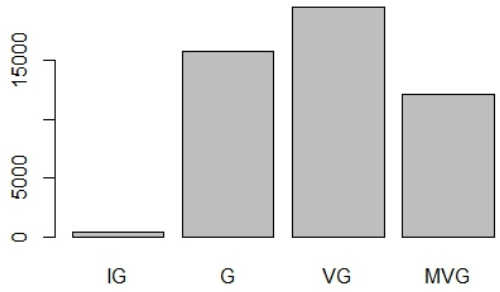
Första matematikkursen i gymnasiet tillhör de 'Gymnasiegemensamma ämnena' och är således en obligatorisk kurs som alla elever skall läsa. Eftersom eleverna läser något olika program läser de också något olika matematikkurser, dock är kursinnehållet såpass likartat att vi väljer att slå ihop alla dessa kurser till endast variabel som representerar elevens betyg i den första matematikkursen. Det finns endast en observation som har betyg "F". Enligt Skolverket skall det inte gå att få en gymnasieexamen utan att få godkänt i denna kursen och därför kommer observationen att tas bort. Se figur 12 för betygsfördelning på kursen.



Figur 12: Betyg Matematik 1, gymnasiet

2.2.8 Matematikbetyg, grundskolan

Om eleven tidigare har haft ett bra betyg i matematik kan detta redan visa att eleven antingen har lätt för att lära sig, har lätt för att förstå matematik eller är motiverad till att studera. Detta bör rimligtvis ha en inverkan för studieresultatet. 376 elever saknar ett inrapporterat betyg för matematik i grundskolan. Förutom den vanliga skalan från IG, G, VG och MVG kan eleven också få ett streck eller bara 'blankt' i sitt betyg. Olika lärare tolkar betygsättningen olika, men vi har i denna studie valt att gå på skolverkets definition som säger att om underlag saknas för en bedömning, så kan ett streck sättas. Vi kommer alltså likställa ett streck och blankt betyg med ett bortfall eftersom vi inte kan yttra oss om elevens kunskaper i ämnet. Det finns 376 personer som saknar ett betyg i matematik från grundskolan. Av dessa saknar också alla 376 ett slutbetyg från åk 9, dvs. mängden av de som saknar matematikbetyg i åk 9 saknar alla även ett slutbetyg i åk 9. Detta fenomen är också rimligt - om en elev inte är godkänt i ett kärnämne kan denne inte få ett slutbetyg. Se figur 13 för histogram beskrivande betygsfördelningen.

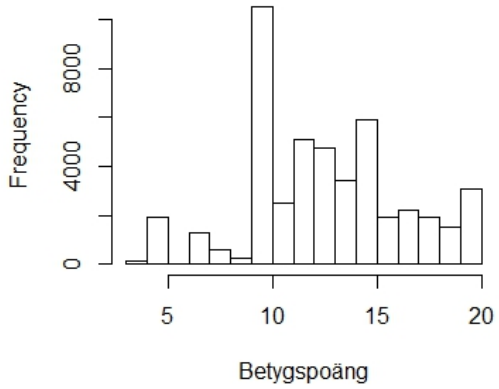


Figur 13: Betyg Matematik, Grundskola

2.2.9 Genomsnittligt matematikbetyg kurs 1, 2 & 3, gymnasiet

Namn	Typ	Beskrivning
$mat123$	Numerisk	$6.67 - 20$

Ett alternativ till att endast ha första matematikbetyget hade varit att ha ett snittbetyg för exempelvis de tre första kurserna. Variabeln är lättare att tolka numeriskt eftersom den får flera möjliga utfall än om man bara har betygsvärdet för en kurs. Ett litet problem som vi är medvetna om är att det ej går att identifiera alla personer som fått icke-godkänt på en kurs eller fler. Vidare kan vi ej heller separera de som bara läst en eller två kurser från de som läst tre. Fördelningen kan ses i figur 14 nedan.



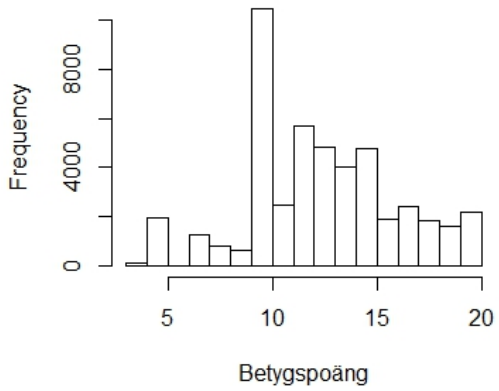
Figur 14: Genomsnittligt betyg i kurs Matematikkurs 1, 2 & 3, gymnasiet

2.2.10 Genomsnittligt matematikbetyg, gymnasiet

Namn	Typ	Beskrivning
$matsnitt$	Numerisk	$3.33 - 20$

Denna variabel och dess fördelning liknar ovanstående väldigt mycket. Vi tar här det genomsnittliga betyget för alla matematikkurser som eleven läst på gymnasiet. Det blir något mer kontinuerliga värden eftersom variabeln kan anta fler värden. Variabeln innehåller även lite mer information än än $mat123$, men möter samma utmaning; en person som läst en kurs och fått ett C i betyg i denna får 15 i genomsnittligt poäng, så även en person som läst två kurser och fått E i ena och A i den andra. Variabeln skiljer alltså inte på dessa, även om det egentligen är olika

utfall. Detta kan eventuellt vara negativt, och är åtminstone värt att notera. Se figur 15 för fördelning.

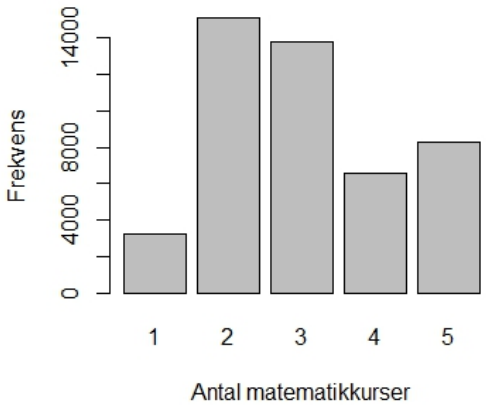


Figur 15: Genomsnittligt betyg i alla Matematikkurser, gymnasiet

2.2.11 Antal matematikkurser, gymnasiet

Namn	Typ	Beskrivning
matkurser	Numerisk	Anger antal matematikkurser eleven har läst i gymnasiet, 1-5

Anger hur många matematikkurser eleven har läst i gymnasiet. Det är obligatoriskt att läsa kurs 1 för att kunna ta ut en gymnasieexamen, dock kan eleven sedan välja att läsa fler kurser. De olika kurserna kan behövas för att få behörighet att söka in till olika universitetsutbildningar, något som är intressant eftersom alla våra program i undersökningen är högskoleförberedande och således kan antas att eleverna är intresserade av någon form av vidareutbildning. Kurs två behövs för en del olika högskoleutbildningar, kurs 3 behövs för att söka inpå merparten av ekonimiprogrammen på universitetet och kurs 4 behövs för att söka tekniska, naturvetenskapliga och ingenjörsprogrammen på högskolenivå. Kurs 5 behövs i regel inte för att komma in på något universitetsprogram, utan är mer en introduktion till högskolematematik och kan väljas till frivilligt. Ibland är kursen obligatoriskt inkluderat i vissa naturvetenskapliga/tekniska program, då eftersom skolan valt detta.



Figur 16: Antal Matematikkurser, gymnasiet

2.2.12 Kommungrupp

Sveriges Kommuner och Landsting har delat in alla Sveriges kommuner i olika grupper baserat på vissa egenskaper. Dessa kan beskådas i nedanstående tabell 3. För utförligare beskrivning och definitioner se referens SKL [3].

Kommungrupp	Benämning	Frekvens
1	Storstäder	3
2	Pendlingskommun nära storstad	43
3	Större städer	21
4	Pendlingskommun nära större städer	52
5	Lågpendlingskommun nära större stad	35
6	Mindre stad/tätort	29
7	Pendlingskommun nära mindre stad/tätort	52
8	Landsbygdkommun	40
9	Landsbygdskommun medbesöksnäring	15

Tabell 3: SKL indelning av kommungrupper

För att göra processen enkel och överskådlig kommer vi att slå ihop kommungrupp nummer ett och nummer tre, se tabell 4 nedan.

Kommungrupp	Benämning	Frekvens	Antal elever
1	Storstäder	24	26 532
0	Övriga kommuner	266	21 790

Tabell 4: Ny indlening av kommungrupper

Vår variabel indikerar nu alltså att kommunen där eleven studerar har ett invånarantal på över 40 000 personer. Som ses faller en majoritet (56 %) av eleverna in under ovanstående kategori.

3 Teori

I det här avsnittet kommer vi kortfattat diskutera teoretisk bakgrund till modeller och metoder som vi kommer använda för att skapa de olika modellerna.

3.1 Regressionsmodeller

3.1.1 Linjär Regression

I praktiken fungerar multipel linjär regression som så att vi löser ett överbestämt linjärt ekvationssystem, dvs. ett ekvationssystem med fler ekvationer än våra okända parametrar. Vi tänker oss att vår responsvariabel *slutbetyg* som vi benämner $\boldsymbol{y}$ kan skrivas som en linjärkombination av observationerna $\boldsymbol{x}$ med parametrarna $\boldsymbol{\beta}$ som koefficienter, samt $\boldsymbol{\epsilon}$ som utgör vår felterm, enligt;

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_k x_{i,k} + \epsilon_i$$

där y_i är observation i av vår responsvariabel, x_{ij} är observation i för förklaringsvariabel j och β_j är parametern associerad till förklaringsvariabel j för totalt n observationer och k parametrar. ϵ_i är den s.k. feltermen för observation i. Vi kommer definiera och diskutera denna mer utförligt under våra antaganden. Under dessa förutsättningar kan vi skriva det överbestämda ekvationssystemet på matrisform med utseendet;

$$\boldsymbol{y} = \boldsymbol{x\beta} + \boldsymbol{\epsilon}$$

där

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \boldsymbol{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & a_{1,k} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,k} \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} \qquad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Eftersom vi har ett överbestämt ekvationssystem där antalet utfall alltså är större än antalet parametrar ($n > k$).

3.1.2 Antaganden

I detta avsnitt kommer vi att gå igenom de viktigaste antagandena som görs för att kunna använda sig av multipel linjär regression. Referenser till dessa hämtar vi från Andersson & Tyrcha (2012) [1].
Det första antagande vi gör är att vår s.k. felterm $\boldsymbol{\epsilon}$ är en oberoende och likafördelad stokastisk variabel (hädanefter förkortar vi detta med s.v.). Detta leder också till att vår responsvariabel $\boldsymbol{Y}$ ses som en s.v. eftersom $\boldsymbol{Y} = \boldsymbol{x\beta} + \boldsymbol{\epsilon}$ och summan av en kostant och en s.v. alltid är en s.v.. Vi ser våra utfall x_{ij} som observerade konstanter utan slump.

Nästa antagande vi gör (som vi nämnde under förra avsnittet) är att vår responsvariabel är en linjärkombination av förklaringsvariablerna och att den kan uttryckas mha. linjära parametrar. Detta gör som sagt att vi kan skriva ekvationerna på formen i förra avsnittet.
Vi kommer också anta att matrisen $\boldsymbol{X}$ har full rang. Detta innebär att ingen av kolonnerna i matrisen utgör en linjärkombination av någon av de andra kolonnerna. Detta är ett viktigt antagande då det behövs för att kunna visa att de lösningar vi hittar till det överbestämda ekvationssystemet har vissa önskvärda egenskaper.
Nu följer ett par antaganden om vår felterm $\boldsymbol{\epsilon}$. Vi antar att feltermen har ett väntevärde lika med noll;

$$\mathrm{E}[\epsilon_i] = 0 \qquad \text{för alla } \; i = 1, 2, ..., n$$

Nästa antagande är att feltermerna har en konstant varians, dvs. värdet av vår observation $x_{i,\cdot}$ skall inte påverka feltermernas varians. Denna egenskap kallas ”homoskedasticitet” och är viktig för flera önskvärda egenskaper när vi utför minskakvadrat-skattningarna för våra β_j.
Vi får alltså

$$\mathrm{Var}[\epsilon_i] = \sigma^2 \qquad \text{för alla } \; i = 1, 2, ..., n$$

Vidare skall det för varje par av unika feltermer inte finnas någon kovarians mellan dessa, dvs. de skall inte vara någon linjärt beroende mellan några två unika feltermer. Detta antagande benämns att vi ej skall ha någon autokorrelation och definieras på så sätt att

$$\mathrm{Cov}[\epsilon_i, \epsilon_j] = 0 \qquad \text{för alla } \; i \neq j$$

Vårt sista antagandet för feltermen $\boldsymbol{\epsilon}$ är att denna är normalfördelad. Detta tillsammans med tidigare antaganden ger då

$$\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

där N är en multivariat normalfördelning, se Gut (2009, s. 120)[10].

3.1.3 β-skattningar

Då vi nu har ett överbestämt linjärt ekvationssystem kommer vi inte kunna hitta en exakt lösning till detta. Vi kommer dock använda oss av en metod som kallas Minsta Kvadrat-metoden (MKM-metoden), vilket går ut på att vi vill hitta skattningarna $\hat{\boldsymbol{\beta}}$ som minimerar summan av kvadraterna av våra feltermer ϵ_i, eller ekvivalent $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$.
Det kommer visa sig att skattningen av $\boldsymbol{\beta}$ som minimerar ovanstående uttryck fås enligt formeln

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

Vi kommer inte härleda detta utan hänvisar istället till Andersson & Tyrcha (2012, s. 20)[1]. MKM-metoden som används för skattningen ger oss flertalet mycket önskvärda egenskaper, bl.a. så är skattningen $\hat{\boldsymbol{\beta}}$ både väntevärdesriktig och

dessutom även den Bästa Linjära Oberoende Estimatorn (BLUE = Best Linear Unbiased Estimator). Detta innebär att den har lägst varians av alla väntevärdesriktiga skattningar, och denna egenskap används då vi skapar konfidensintervall för $\hat{\boldsymbol{\beta}}$. Återigen hänvisas till Andersson & Tyrcha (2012)[1] för mer utförlig Teori.

3.1.4 Odds och Oddskvot

En del av våra variabler utgörs av kategorisk data. Det är data med utfall som klassificeras inom olika kategorier såsom; Ja & Nej, Lyckad & Misslyckad eller Röd, Blå & Grön. Ett begrep som ofta används när man talar om kategoridata är *Odds*. Oddset ger ett ratio mellan sannolikeheten att något skall inträffa gentemot sannolikheten att det inte inträffar och definieras som;

$$Odds\,(Y = y) = \frac{P(Y = y)}{1 - P(Y = y)}$$

För att mäta hur proportionerna för att lyckas förändras givet att en (förklarings-)variabel skiftar nivå används oddskvoten. Denna definieras enligt:

$$Oddskvot\,(Y = y|X = x_1) = \frac{Odds\,(Y = y|X = x_1)}{Odds\,(Y = y|X = x_0)}$$

I detta fall mäter vi hur sannolikheten för att utfall y inträffar givet att x_0 skiftar nivå till x_1. Denna förändring gäller alltid i jäfmörelse med en på förhand vald basnivå; det är den nivån som förändringen i utfallet jämförs med. I formeln ovan är x_0 basnivån. I fallet då X är en numerisk variabel ger oddskvoten ett mått på hur oddset förändras per ökad enhet i X, t.ex. per +1 poäng eller per +1 cm.

3.1.5 Logistisk regression

Med hjälp av de definitionerna från förra avsnittet kan vi definiera en ny typ av regressionsmodell, nämligen den logistiska regressionsmodellen;

$$\log[Odds\,(Y = y|\boldsymbol{X}] = \boldsymbol{x})] = \alpha + \sum_{i=1}^{p} \beta_{x_i}^{X_i}$$

där $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$, $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$ och $\beta_{x_i}^{X_i}$ är den skattade parametern för varibalen X_i då denna antar nivå x_i. I det fallet då X_i är en kontinuerlig variabel är $\beta_{x_i}^{X_i} = \beta^{X_i} \cdot x_i$ istället. Modellen lämpar sig väl för kategoriska responsvariabler, speciellt binära sådana. Med detta menas att responsvariabeln är på sådan form att antingen så inträffar händelsen $Y = y$, eller så inträffar den inte. Vid flera utfallskategorier är modellen inte lika lämplig eftersom den endast beskriver proportionen mellan två olika händelser. Modellen är lätt att arbeta med eftersom oddskvoterna för alla förklarande variabler är enkla att beräkna från parameterskattningarna då $Oddskvot(Y = y|X_i = x_i) = e^{\beta_{x_i}^{X_i}}$. Iteratively reweighted least square-metoden används vanligen för parametrarna i den logistiska regressionsmodellen och även så i programvaran vi använder [8].

3.2 Selektion & Verifiering av Modeller

Hur vet man om en modell är bra? Förutom att antagandena måste vara uppfyllda finns det en mängd olika verktyg och mätvärden att se på. Här kommer vi förklara, definiera och troliggöra några av de vanligaste metoderna.

3.2.1 R^2 och justerat R^2

I denna avdelning använder definitioner från Andersson & Tyrcha (2012)[1]. När vi väljer en linjär modell finns det flera metoder att använda sig av och ta hänsyn till. R^2 värdet ger oss ett mått på hur väl vår modell förklarar resultatet.
För att definiera R^2 behöver vi dock studera några uttryck som förekommer i vår regression. Vi börjar med att definiera "Total Sum of Squares" (TSS). Detta kan alltså tolkas som summan av kvadraterna av de totala avvikelserna i de faktiska utfallen mot vår modell. I matrisform ser uttrycket ut på följande sätt;

$$TSS = \boldsymbol{Y}^T\boldsymbol{Y} - n\bar{Y}^2$$

Vi definierar nu ytterligare två uttryck, nämligen "Explained Sum of Squares" (ESS) samt "Residual Sum of Squares" (RSS). I matrisform ser uttrycken ut på följande sätt;

$$ESS = \hat{\boldsymbol{\beta}}\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} - n\bar{Y}^2, \quad RSS = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$$

Sambandet $TSS = ESS + RSS$ gäller. ESS kan tolkas som summan av de kvadrerade avvikelserna från vår skattning gentemot medelvärdet av utfallen, emedan RSS är summan över de kvadrerade felen i vår modell. Vi har alltså delat upp avvikelserna från utfall i en term som vår modell förklarar, samt en term som vår modell inte kan förklara, dvs. "felet" ($= \boldsymbol{\epsilon}$). För fullständiga bevis och härledningar hänvisar vi till Andersson & Tyrcha (2012, s. 29)[1].
Nu är vi redo att definiera R^2-värdet som är ett mått på hur mycket av avikkelserna vår modell kan förklara, dvs. andelen av vår förklarande kvadratsumma kontra den totala kvadratsumman; $R^2 = \frac{ESS}{TSS}$. Vi kommer dock använda ett justerat R^2-värde som vi kallar för R^2_{adj} eftersom detta mätvärde också tar hänsyn till antalet parametrar i modellen.

$$R^2_{adj} = 1 - \frac{ESS/(n-k)}{TSS/(n-1)}$$

3.2.2 Goodness of Fit

Goodness of Fit för en statistisk modell beskriver testar hur väl modellen kan prediktera datan i fråga. Det finns många olika test för att undersöka hur väl en regressionsmodell predikterar ett visst datamaterial. För en binär logistisk regressionsmodell finns exempelvis Pearson's χ^2-test eller Likelihood-ratio G^2 statistika. Enligt Agresti (2013, s. 172)[5] konvergerar dock inte dessa två statistikor i det fall då datamaterialet innehåller ogrupperad data, nästan-kontinuerliga variabler eller kontinuerliga variabler, vilket gör att vi kommer behöva använda ett Hosmer-Lemeshow test för att utvärdera vår Goodness of Fit för vår logistiska regression.

Testet börjar med att skatta sannolikheten att lyckas under modellen i fråga. Sedan delas datamaterialet in i ett antal mindre grupper, ofta tio stycken. Grupperna skapas baserat på de skattade sannolikheterna och sorteras i stigande ordning. Därefter jämförs de skattade sannolikheterna med de observerade utfallen inom respektive grupp för att se hur väl dessa överensstämmer.

$$G_{HL} = \sum_{i=1}^{g} \frac{(\sum_j y_{ij} - \sum_j \hat{\pi}_{ij})^2}{(\sum_j \hat{\pi}_{ij})[1 - (\sum_j \hat{\pi}_{ij})/n_i]}$$

Där y_{ij} är det observerade utfallet (1 eller 0), $\hat{\pi}_{ij}$ är den skattade sannolikheten för observation $j = 1, 2, \ldots, n_i$ i grupp $i = 1, 2, \ldots, g$ där g är det totala antalet subgrupper som populationen delats in i.
G_{HL}-testet är approximativt χ^2_{g-2}-fördelat. Definitionen är enligt Agresti (2013, s. 173)[5].

3.2.3 Akaike informationskriterium

Ett annat värde som används för att utvärdera modeller är Akaike informationskriterium som vi kommer benämna AIC. Vi kommer därför också se på AIC-värdet, vilket också ger oss ett mått på modellens förklaringsgrad dels mha. log-likelihood-funktionen, men värdet innehåller också en del som bestraffar modeller med många variabler, dvs. en modell med färre förklaringsvariabler som har samma log-likelihood-värde kommer få ett högre AIC-värde och anses bättre. Det betyder att vi på så vis kan undvika överanpssning, vilket passar oss bra eftersom vi strävar efter att förklara resultatet med en lättolkad och så enkel modell som möjligt. För definitionen använder vi oss av de definitioner som ges av Agresti (2013, s. 212)[5].
AIC poängsätter en modell mha. av formeln

$$AIC = -2(\log(\mathcal{L}) - k)$$

Där $\mathcal{L}$ är det maximala värdet av Likelihood-funktionen för modellen och k är antalet parametrar i modellen.

3.2.4 Stepwise Regression

Vi kommer använda oss av funktionen `step()` som alltså är stepwise-regression. `Step()` beskrivs av Sundberg (2014)[6] och fungerar på så sätt att vi börjar med en tom modell som endast innehåller en konstant. Sedan testas alla variabler i modellen för att se vilket AIC-värde som skulle erhållas då respektive variabel lades till, och funktionen väljer slutligen ut den som ger bäst resultat. Efter detta ser funktionen efter om den skulle kunna minska AIC-värdet genom att plocka bort någon av variablerna, och om så är fallet görs detta. Sedan upprepas proceduren. Funktionen kan också välja ut förklaringsvariabler baserat på p-värdet; när variabeln kollas så gör funktionen ett test för att se efter vilken variabel som ger det mest signifikanta utslaget (dvs. har minst p-värde) i testet $H_0 : \beta_i = 0$. Efter varje variabel som vi lägger till i modellen testas sedan alla befintliga parametrar i modellen var för sig och vi kontrollerar då så att den nya variabeln inte har ändrat de gamla variablernas signifikans.

3.2.5 Variations Inflationsfaktor

"Variance Inflation Factor" (här kallat VIF) är ett mätvärde som kan användas för att se i hur stor grad de olika förklaringsvariablerna är linjärkombinationer av varandra (dvs. om de innehåller samma information). Definitionen av VIF för en linjär regressionsmodell är

$$VIF = \frac{1}{1 - R_j^2}$$

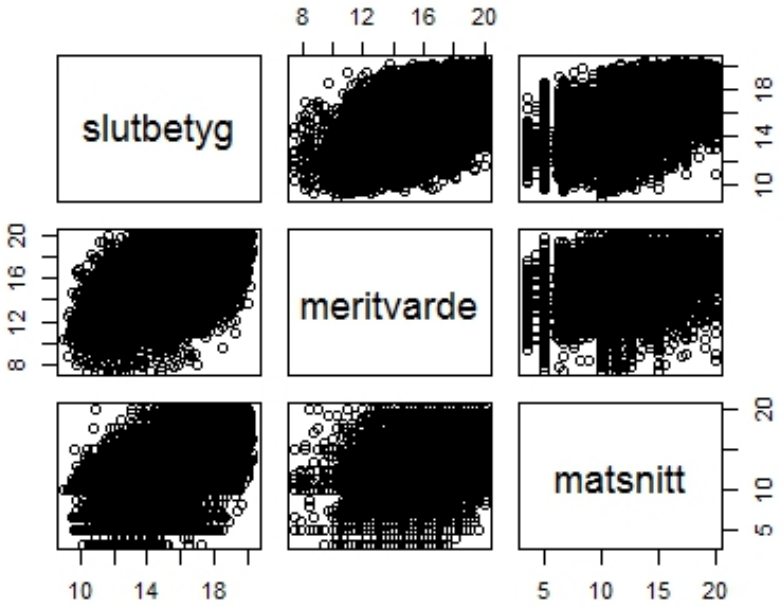
Där R_j^2 står för den förklaringsgrad som anger hur mycket av variationen i x_j som förklaras av de andra x-variablerna. Som tumregel kan man säga att ett högt VIF-värde är runt 5 och över, detta enl. Sundberg (2014, s. 73)[6], och det kan då vara läge att undersöka sina förklaringsvariabler noggrannare för att se om det föreligger multikollinjäritet.
I fallet med en logistisk regression måste vi använda oss av den generella metoden som kallas "Generalized Variance Inflation Factor" och ges Fox & Monette (1992, s. 179)[7]. Jag måste erkänna att jag inte riktigt är i stånd att förklara denna definition på ett övergripligt sätt, således lämnar vi endast referensen och läsaren kan själv ta del av den utförligare definitionen och härledningen.

4 Analys

I den här avdelningen kommer vi att testa två stycken regressionsmodeller, en linjär och en logistisk modell. Vi använder oss av statistik-programmet R (Version 0.98.1103). Efter databehandlingen i kapitel två utgörs vår population av 47028 fullständiga utfall och det är dessa som vi alltså kommer att utföra analysen på.

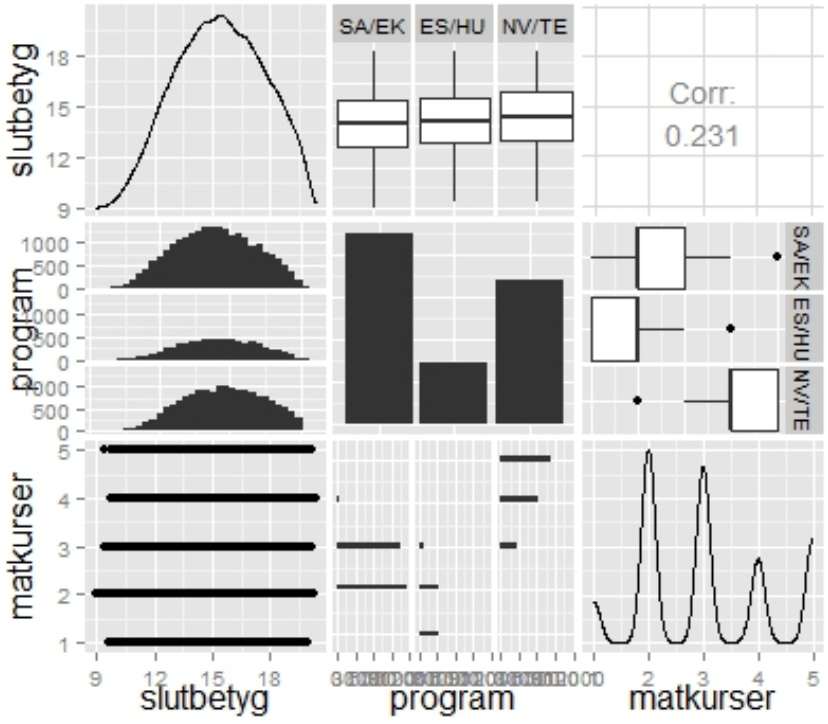
4.1 Korrelationsmatriser

För att få en överblick av hur våra förklaringsvariabler kan komma att samspela med varandra och mot vår responsvariabel skall vi plotta dessa i ett par matriser. Utifrån matriserna ska vi försöka läsa av några samspel och även se över våra antaganden för de statistiska modellerna vi ämnar använda oss av. Vi kommer främst titta på de numeriska variablerna eftersom det kan vara lättare att se samband i dessa. I vår första matris 17 kan vi se att det är rimligt att anta att det råder ett linjärt samband mellan förklaringsvariablerna och responsvariabeln. Vi kan också se att de båda förklaringsvariablerna visar vissa tecken på linjärt samband; korrelationen (Pearson) mellan *matsnitt* och *meritvarde* är 0.61. Detta är ganska rimligt och korrelationen ser inte ut att vara så pass hög att det kommer ställa till problem i form av att vi bryter mot antagande om linjärt oberoende mellan förklaringsvariablerna.



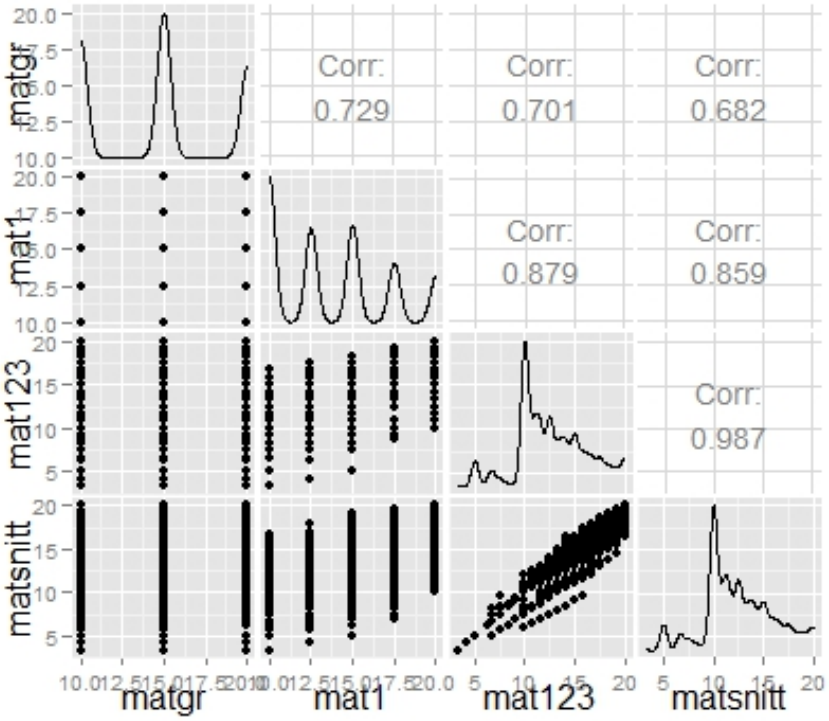
Figur 17: Respons, meritvärde och genomsnittligt matematikbetyg i gymnasiet

I figuren 18 kan vi se hur variablerna samvarierar. Det är tydligt att antal matematikkurser är klart överrepresenterade i naturvetenskap- och teknikprogrammen och de som läser Estet- eller Humanistprogrammet läser klart minst matematik. Vidare ser det ut att finnas ett svagt positivt samband mellan slutbetyget och att läsa fler matematikkurser då korrelationen mellan dessa är 0.23, något som dock inte går att urskilja baserat på figuren längst ner i vänstra hörnet. På diagonalen kan fördelningen ses för respektive variabel.



Figur 18: Respons, program samt antal matematikkurser

I den sista figuren 19 jämför vi de olika variablerna som vi har tagit fram för att mäta matematikbetyget. Alla dessa har väldigt hög korrelation sinsemellan, vilket är rimligt eftersom de mäter mer eller mindre samma sak. I modellen vill vi gärna inkludera med matematikbetyget från åk 9 samt en av de olika variablerna som representerar matematikbetyget i gymnasiet. Vi kan se att den som passar bäst för detta ändamål är *matsnitt*, då dessa två är de som har lägst korrelation sinsemellan (0.68), och därför kommer denna variabel väljas in i den första modellen. Återigen finns respektive fördelnig för de enskilda variablerna utritade på diagonalen.



Figur 19: Matematikbetygsvariabler

4.2 Linjär regression

Vårt mål är att på ett enkelt och tolkningsvänligt sätt försöka förklara en elevers slutbetyg utan att förlora alltför mycket i förklaringsgrad (R^2_{adj}-värdet). I första delen kommer vi att härleda i vår analys steg för steg. I andra delen utför vi ett antal test för att se om vår modell håller.

Vi börjar med att göra en backwards-elimination med en helt komplett modell. I detta steget tas inga variabler bort. En snabb test med stepwise regression där vi börjar med en tom modell så väljer programmet att fylla upp till den fullskaliga modellen, dvs. Samma modell som backwards-elimination gav oss. Fullständiga resultat och utskrifter från R för den första modellen med dess olika variabelskattningar, AIC- och VIF-värden kan ses i Appendix.

Vi noterar att den första modellen har en förklaringsgrad med ett R^2_{adj}-värde på 0.66. I denna modell har alla variabler valts med, vilket inte är optimalt eftersom en del variabler inte är signifikanta, samt att vissa av dessa innehåller snarlik information. Vi börjar med att kolinjaritet mha. VIF-värdena; två variabler har något högt VIF-värde, nämligen *matkurser* och *program*, båda med runt 5 i värde. Detta innebär att vi misstänker att viss kolinjaritet kan föreligga. Matematikbetyget i grundskolan har ett vif-värde på 2.7, och de övriga variablerna har alla väldigt låga VIF-värden på runt $1 - 1.2$. Vi noterar alla dessa VIF-värden men väljer att inte åtgärda något ännu eftersom vi saknar signifikans i vissa av våra variabler.

Som sagt är några av våra kategoriska variabler och deras utfall är inte signifikanta, nämligen $program = SA$ och $foruniva = 2$. Vi börjar med att slå ihop programmen enligt avdelningen i kapitel 2, databehandling. Då vi strävar efter att ha en enkel och tolkningsbar modell bidrar denna sammanslagning till att göra modellen mer lättöversiktlig. Se kapitel 2 för vidare motivering och mer utförlig beskrivning och fördelning.
Föräldrarnas utbildningsnivå slår vi ihop kategori 1 med kategori 2, detta betyder att basfallet och den lägsta utbildningsnivån för den högst utbildade föräldern i familjen nu är en gymnasieexamen. Eftersom de olika betaskattningarna för utländsk bakgrund är svårtolkade och inte riktigt ser rimliga ut, slår vi samman dessa. Vi slår alltså samman grupp 2 med grupp 1 så att variabeln nu endast indikerar om det finns utländsk bakgrund eller ej (basfallet $utlbakg = 0$, ingen utländsk bakgrund). För utförligare beskrivning hänvisar vi till kapitel 2, databehandling. Vi kommer även motivera denna sammanslagning mer utförligt i diskussionen.

Vi kör step-funktionen återigen som ännu en gång väljer att inte exkludera någon av förklaringsvariablerna. Vi behåller vårt R^2_{adj}-värde på 0.66. Vi ser över våra VIF-värden och ser att *matkurser* och *program* fortfarande har relativt höra värden på 4.5 respektive 3.9. Viss kolinjaritet kan alltså misstänkas att föreligga vilket låter rimligt - natur- och teknikprogramstudenter läser i mycket högre grad matematik än vad de övriga programmen läser, och båda variablerna har fått negativt skattade koefficienter. Vi utesluter *matkurser* då denna har högst VIF-värde samt att det är en rimligare tolkning i att säga att det är bestraffande för slutbetyget att läsa ett visst program än att äga att det är negativt att läsa många matematikkurser. När en elev söker till gymnasieskolan är dock det första valet man gör vilket program man skall gå, inte hur många matematikkurser man skall läsa. Därför är det naturligare att låta *program* vara den förklarande variabeln, eftersom det är detta val ligger eleverna närmare till hands och om man väljer estetprogrammet väljer man det sannolikt delvis för att slippa läsa matematikkurser.

Variabelskattningarna i modellen ser i en första anblick alla rimliga ut, sånär som på en; nämligen $matgr = -0.072$. Denna är negativ och visar alltså på att för att uppnå ett gott slutbetyg i gymnasiet är det en nackdel att ha ett högt matematikbetyg i grundskolan. Mer bestämt tolkas det som att ju bättre matematikbetyg som en elev har i åk 9, desto sämre slutbetyg förväntas denna få i gymnasiet. Gissningsvis ser vi en samverkanseffekt mellan denna variabel och någon av de andra,

alternativt går det att spekulera i någon form av betygsinflation i grundskolan. Variabeln har ett VIF-värde på 2.64, vilket inte är anmärkningsvärt högt i sig. Det är dock i nuläget det högsta VIF-värdet i modellen och eftersom vi har ett negativt samband till slutbetyget provar och utesluta variabeln. Vi ser då lägre VIF-värden för alla övriga variabler (som nu alla ligger under 2) samt en ytterst marginell minskning i förklaringsgrad på endast 0.0056, så avrundat står vi ändå kvar på 0.66. Eftersom vi vill förklara slutbetyget på ett enkelt och intuitivt sätt som möjligt utesluter vi denna variabel.

Vi landar nu i vår slutgiltiga modell för den multipla linjära regressionen, och dess skattade koefficienter kan beskådas i tabellen nedan.

	Estimat	Std. avvikelse
α	5.38	0.044
$meritvarde$	0.43	0.003
$utlbakg = 1$	-0.18	0.018
$foruniva = 3$	0.06	0.017
$foruniva = 4$	0.15	0.015
$kon = 1$	-0.33	0.013
$storstad = 1$	-0.13	0.015
$friskola = 1$	0.42	0.014
$program = ES/HU$	0.17	0.018
$program = NV/TE$	-0.78	0.015
$studieomf = 1$	0.44	0.014
$matsnitt$	0.28	0.002

4.2.1 Exempel

Vi illustrerar hur modellen fungerar nedan med mig själv som ett fiktivt exempel. Vi säger att jag att jag gick naturprogrammet på en kommunal skola i ett storstadsområde och läste utökat antal kurser. Den högst utbildade föräldern hade en högskoleutbildning som var längre än tre år, och enligt specificerade kategorierna finns inget utländskt påbrå. Från grundskolan hade jag ett meritvärde på 17.5 poäng och jag hade ett genomsnittligt betyg i matematik på 15 i gymnasiet. Jag är man. Modellen ger oss för dessa data;

$$\begin{aligned} slutbetyg &= \alpha + \beta_{\boldsymbol{x}}^{\boldsymbol{X}} \\ &= \alpha + \beta^{meritvarde} \cdot meritvarde + \beta^{program} + \beta^{matsnitt} \cdot matsnitt \\ &+ \beta^{studieomf} + \beta^{kon} + \beta^{foruniva} + \beta^{utlbakg} + \beta^{storstad} + \beta^{friskola} \\ &= 5.38 + 0.43 \cdot 17.5 - 0.78 + 0.28 \cdot 15 + 0.44 - 0.30 + 0.15 + 0 - 0.13 + 0 \\ &= 16.49 \end{aligned}$$

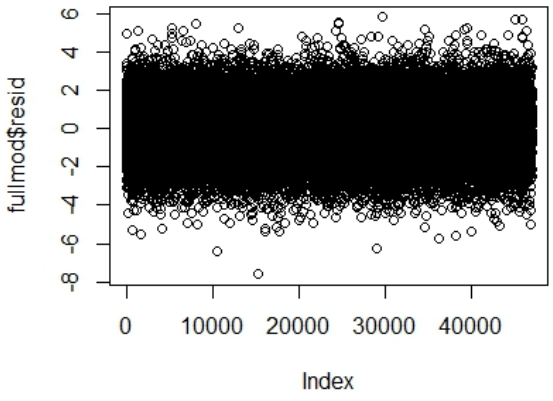
Mitt förväntade slutbetyget för följande värden på förklaringsvariablerna är alltså 16.49, enligt modellen.

4.2.2 Outliers

`outlierTest()` är en funktion som ser över våra utfall, dvs. y_i under antagande att modellen som vi testar är sann. Testet påvisar att vi har en observation, som under modellen, har ett p-värde (Bonferroni) på ca 0.0004; feltermen är hela -5.74. Vi undersöker utfallet närmare och ser att eleven har högt meritvärde från grundskolan (16.94), välutbildade föräldrar (kategori 4), A i båda matematikkurser denne läst i gymnasiet samt att eleven är ambitiös och läser utökade kurser. Eleven gick estetprogrammet men går endast ut med 10.5 i slutbetyg. Vi noterar denna förekomst men kommer inte vidta åtgärder eller misstänka något fel pga. detta; att ett exempel har ett sådant extremt utfall kan ses som ett fall av varians och observationen blir inte så inflytelserik då vi har en stor population.

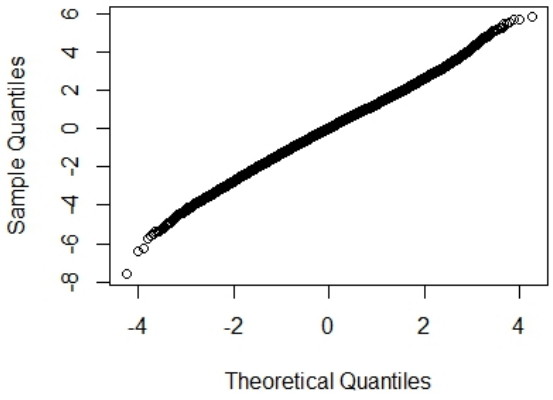
4.2.3 Residualer

I detta avsnitt har vi för avsikt att se efter att våra antaganden om våra feltermer ϵ_j stämmer. De predikterade värdena $\hat{Y}$ av Modell 1 stämmer inte helt överens med Y (detta vet vi eftersom $R^2 \neq 1$), utan vi får ett väntat "fel" i modellen - den slump som vi inte kan förklara. Differensen $Y_i - \hat{Y}_i$ kallar vi för residualer. Dessa kan ses som en realisering av vår felterm ϵ_j, och genom att studera deras fördelning kan vi se om våra antaganden om dem verkar stämma eller ej. Vi börjar med att plotta residualerna i figur 20.



Figur 20: Residualplot

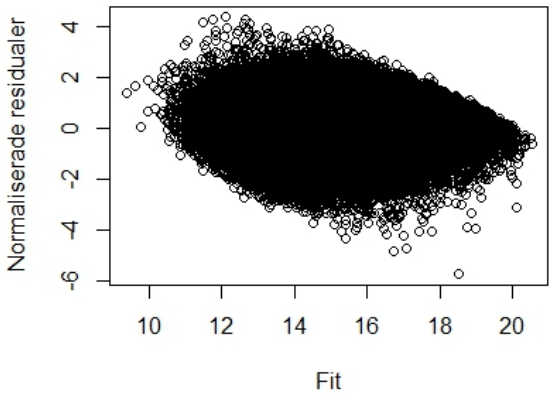
Vi ser inga konstigheter i denna eftersom det inte finns något tydligt mönster i spridningen - våra antaganden om att feltermerna är oberoende och likafördelade ser ut att hålla.



Figur 21: Approximativt Nomralfördelade Residualer

I figur 21 jämför vi kvantilerna på feltermerna med kvantilerna på en normalfördelning. Vi ser att residualerna till en stor del ligger på linjen och kan därför antas vara approximativt normalfördelade vilket också var ett av våra antaganden, vi finner alltså inga belägg för att påvisa att vi bryter mot detta antagande.

I den sista figuren 22 har vi plottat våra residualer mot våra skattade värden $\hat{y}$ (Fit).



Figur 22: Normaliserade residualer plottade mot $\hat{y}$

Här noterar vi att det eventuellt kan föreligga heteroskedasticitet, dvs. vi bryter eventuellt mot antagande om homoskedasticitet. Detta misstänker vi eftersom det syns en svagt nedåtsluttande "kam" i övre högra delen av vår plot 22. Eftersom vi har ett positivt linjärt samband håller inte modellen för de som har ett, enligt modellen, förväntat gott betyg; vi vill ha en jämn spridning av feltermerna runt

det förväntade värdet, men eftersom det inte går att få ett högre betyg än 20.0 i slutbetyg blir således felen, givet goda in-värden på förklaringsvariablerna, oundvikligen negativa i större utsträckning. Detta betyder mao. att vårt anatagande om homoskedasticitet inte ser ut att hålla.

En av de mer dominerande termerna i vår prediktion är *meritvarde* och vi misstänker att om vi kan rätta till residualerna i förhållande till *meritvarde* så kan eventuellt också residualerna uppföra sig på ett önskvärt sätt jämfört med $\hat{Y}$. Vi försöker även transformera responsvariabeln på några olika sätt, bla. $e^{slutbetyg}$, $log(slutbetyg)$ samt $\frac{1}{1+slutbetyg}$. Vi provar att göra samtliga nämnda transformationer på vår responsvariabler men ingen av dessa åtgärder ger oss det resultatet vi önskar - residualerna visar samma tecken på heteroskedasticitet.

När våra antaganden inte håller för modellen innebär detta tyvärr att vi inte kan tillskriva vår modell de önskvärda statistiska egenskaper som Minsta Kvadratskattningsmodetoden medför; t.ex. vet vi nu inte att våra β-skattningar är de Bästa Linjära Oberoende Estimatorerna (BLUE=Best Linear Unbiased Estimator) se Andersson & Tyrcha (2012)[1]. Vidare riskerar också konfidensintervallet för $\hat{\beta}_i$ att inte bli korrekt då vi kan få fel värden i vår t-fördelning - terortiskt sett kan det mao. finnas en skattare med lägre varians. Eftersom modellantagandena inte är uppfyllda och att detta leder till brister anser vi att modellen inte håller och således ej är lämplig för att förklara elevernas studieresultat i gymnasiet.

4.3 Logistisk regression

I denna modell kommer vi att ha samma variabler som för den linjära modellen, med undantag av att vi kommer ha en kategorisk responsvariabel som anger om en elev har fått ett bra slutbetyg eller ej. Vi har valt att se ett bra slutbetyg som ett genomsnittsbetyg på C eller bättre, vilket innebär att eleven skall ha en genomsnittlig betygspoäng på 15 eller bättre. Vi kallar variabeln för *slutbetyg_overC*. Totalt hade 11513 personer, eller 55.61% av populationen, ett slutbetyg som motsvarade C i genomsnitt eller bättre (dvs. ett slutbetyg på 15 eller bättre).
Vi börjar med att köra stepwise-funktionen med en tom modell, som slutar med att alla variabler väljs in till modellen. Liknande fallet med den linjära modellen saknar vi signifikans i några variabler, och vi får dessutom något höga VIF-värden på *program* (5.2), *matkurser* (4.74) och *matgr* (2.14), men vi vidtar inga åtgärder mot detta ännu. För fulltständig utskrift av koefficienter, stepwise-regression och VIF-värden från första modellen, se appendix.
Vidare ser vi att det saknas signifikanser i vissa variabler och/eller utfallskategorier och för att råda bot på detta slår vi samman följande kategorier; $foruniva = 1$ slås ihop med $foruniva = 2$ och $foruniva = 3$ slås ihop med $foruniva = 4$. Tabellen 5 nedan illustrerar.

Gammal kategori	Ny kategori
1	0
2	0
3	1
4	1

Tabell 5: Nyckel till föräldrars utbildningsnivå efter sammanslagning

I praktiken innebär det nu att varibalen $foruniva$ nu har två nivåer där $foruniva = 1$ indikerar att den högst utbildade föräldern har någon form av högskoleutbildning och $foruniva = 0$, som är basfallet, indikerar att ingen av föräldrarna har högskoleutbildning.
Vidare slås $utlbakg = 2$ samman med $utlbakg = 1$ med liknande resonemang som i den linjära multipla regressionen. Variabeln skiljer nu ej på olika generationer av invandring, utan bara om det finns utländsk bakgrund i familjen eller ej, dvs. $utlbakg = 1$ indikerar att båda föräldrar är födda utomlands, eller att eleven själv är född utomlands. *program* slås samman på exakt samma sätt som de gjordes i den linjära regressionsmodellen, se kapitel 2 databehandling för utförlig beskrivning.

Stepwise-funktionen väljer nu ut samtliga variabler igen och denna gång får vi signifikans för alla variabler och kategorier. Vi noterar dock att *matkurser* och *program* har höga VIF-värden; 4.20 respektive 3.93, varför vi väljer att utesluta den förstnämnda. Analogt med i den linjära analysen måste man ställa sig frågan vad *matkurser* ger oss för information och vi skall diskutera mer om detta i Resultat-delen.
Liknande den linjära modellen har vi en variabelskattning som ser konstig ut; $matgr = -0.10$. Tolkningen av detta skulle alltså vara att för varje betygspoäng bättre än 10 (= precis godkänt betyg) i grundskolematematik så skulle oddset

för att få ett genomsnittsbetyg över C försämras med ca 10% (eftersom $e^{-0.10} = 0.90$). Detta låter inte trovärdigt eftersom ett gott betyg inte bör medföra sämre framtidsutsikter för slutbetyget, och vi väljer att utesluta parametern *matgr*. När vi har gjort detta ser vi även här en positiv effekt på VIF-värdena som sjunker för samtliga parametrar, inget värde överstiger 1.5. Efter denna uteslutning får vi följande modell som också blir vår slutgiltiga;

	Estimat	Std. avvikelse
α	-12.53	0.128
$meritvarde$	0.53	0.008
$utlbakg = 1$	-0.24	0.037
$foruniva = 1$	0.15	0.028
$kon = 1$	-0.50	0.028
$storstad = 1$	-0.16	0.031
$friskola = 1$	0.59	0.031
$program = ES/HU$	0.29	0.038
$program = NV/TE$	-1.10	0.032
$studieomf = 1$	0.65	0.030
$matsnitt$	0.40	0.006

För att verifiera modellen använder vi ett Hosmer-Lemeshow goodness of fit-test mha. funktionen `hoslem.test()`, funktion finns som del i paketet `ResourceSelection`. Under detta test är H_0 att den logistiska modellen i fråga inte kan förklara något av utfallen. Vi får dock ett p-värde på under 0.001 ($p \leq 2.2 \cot 10^{-16}$), vilket betyder att vi kan förkasta H_0 på 99.9% signifikansnivå och istället anta att modellen är sann. Nedan ses en tabell med respektive variabels oddskvot.

	Oddskvot
$meritvarde$	1.70
$utlbakg = 1$	0.78
$foruniva = 1$	1.16
$kon = 1$	0.61
$storstad = 1$	0.85
$friskola = 1$	1.80
$program = ES/HU$	1.33
$program = NV/TE$	0.33
$studieomf = 1$	1.91
$matsnitt$	1.49

Tabell 6: Oddksvoter

modellen visar nu att *meritvarde* har ett positivt inflytande, för varje hel poäng meritvärde från åk 9 som studenten besitter (utöver 7.5 poäng som är miniminivå) ökar oddset med 70% för att få ett slutbetyg som är bättre än ett C i genomsnitt. Ett merkant resultat är att det ser ut att finnas stora skillnader mellan vilket program eleven väljer att gå; om eleven går Estet- eller Humanistlinjen ökas oddset för att få ett genomsnittsbetyg över C med 133% gentemot att läsa Ekonomi- eller

Samhällsvetenskapsprogrammet. Om eleven istället går Naturvetenskapligt eller tekniskt program så minskar oddset med hela 67% för att lyckas få ett genomsnittsbetyg på C eller bättre.
Vi testar även att omdefiniera ett bra betyg till att utgöra ett genomsnittligt slutbetyg som lika med eller bättre än ett 'D' i genomsnitt, dvs. ha ett slutbetyg större eller lika med 12.5 poäng. Vi kör den logistiska regressionen igen med samma parametrar och får ett liknande resultat som vår tidigare modell. Några noterbara förändringar är dock att det är mer straffbart att vara man i denna modell, samt att det inte är fullt lika straffbart att läsa ett Naturvetenskapligt eller tekniskt program. Vi gör även en modell där ett bra betyg ses som B eller bättre. Vi kan se att även denna inte skiljer sig anmärkningsvärt mycket mot vår ursprungliga modell. För fullständiga skattningar av parametrar för dessa alternativa modeller, se tabeller i appendix.

4.3.1 Exempel

Precis som i fallet med den multipla linjära regressionsmodellen kommer vi illustrera hur den logistiska modellen fungerar nedan med mig själv som ett fiktivt exempel.
Vi säger att jag att jag gick naturprogrammet på en kommunal skola i ett storstadsområde och läste utökat antal kurser. Den högst utbildade föräldern hade en högskoleutbildning som var längre än tre år, och enligt specificerade kategorierna finns inget utländskt påbrå. Från grundskolan hade jag ett meritvärde på 17.5 poäng och jag hade ett genomsnittligt betyg i matematik på 15 i gymnasiet. Jag är man. Modellen ger oss för dessa data;

$$\begin{aligned}
\log[Odds(Betyg \geq C)] &= \alpha + \beta_{\boldsymbol{x}}^{\boldsymbol{X}} \\
&= \alpha + \beta^{meritvarde} \cdot meritvarde + \beta^{program} + \beta^{matsnitt} \cdot matsnitt \\
&+ \beta^{studieomf} + \beta^{kon} + \beta^{foruniva} + \beta^{utlbakg} + \beta^{storstad} + \beta^{friskola} \\
&= -12.53 + 0.53 \cdot 17.5 - 1.10 + 0.40 \cdot 15 + 0.65 - 0.50 + 0.15 + 0 - 0.16 + 0 \\
&= 1.79
\end{aligned}$$

detta ger oss sedan sannolikheten för att jag skall gå ut gymnasiet med en genomsnittspoäng på C eller bättre enligt

$$P(Betyg \geq C) = \frac{e^{1.79}}{1 + e^{1.79}} = 0.86$$

Under dessa förutsättningar är alltså sannolikheten för att jag skall få ett slutbetyg på 15 poäng eller bättre 86 %.

5 Diskussion

I analysdelen har vi funnit två modeller och gett härledningar till hur vi kommit fram till dessa. Den första modellen, vår multipla linjära regressions-modell, har dock påvisat egenskaper som ger oss starka anledningar att misstänka att modellen inte håller. Av denna anledning kommer resultatdelen främst baseras på den logistiska regressionsmodellen.

5.1 Resultat

I tabell 7 nedan publiceras Oddskvoterna för våra skattade parametrar tillsammans med 95% konfidensintervall för skattningarna.

	Oddskvot	2.5%-kvantil	97.5%-kvantil
$meritvarde$	1.70	1.67	1.72
$utlbakg = 1$	0.78	0.72	0.84
$foruniva = 1$	1.16	1.09	1.22
$kon = 1$	0.61	0.58	0.64
$storstad = 1$	0.85	0.80	0.90
$friskola = 1$	1.80	1.70	1.91
$program = ES/HU$	1.33	1.24	1.43
$program = NV/TE$	0.33	0.31	0.35
$studieomf = 1$	1.91	1.80	2.02
$matsnitt$	1.49	1.47	1.51

Tabell 7: Oddksvoter med 95% konfidensintervall

För få kvantilerna används att $\hat{\beta} \pm z_{\alpha/2}(SE)$, detta enligt Agresti (2012, s. 106)[5], där z är $N(0, 1)$-fördelad och SE står för standardavvikelsen.

5.1.1 Meritvärde i åk. 9

Att det finns ett väldigt starkt samband med elevens slutbetyg i åk 9 och slutbetyget i gymnasiet lär inte ses som någon överraskning. Oddskvoten för meritvärdet visar på att oddset för att få ett bra betyg ökar med 70 % (67, 72) per enhet tillskott i meritvärde (utöver 7.5 vilket är det minsta värdet som meritvärdet i populationen kan anta). Meritvärdet kan ju sägas vara grundskolans motsvarighet mot gymnasiets slutbetyg och det är rimligt att en elev som redan har goda kunskaper från grundskolan har har klart bättre förutsättningar att nå ett gott slutbetyg i gymnasiet.

5.1.2 Utländsk bakgrund

Oddskvoten för utländsk bakgrund säger oss att oddset för att få ett bra betyg reduceras med 22 % (16, 28) då eleven har en utländsk bakgrund; alltså antingen om denne själv är född utomlands eller om båda föräldrarna är födda utomlands.

Det kan anses tänkbart att elever med utländsk bakgrund kan ha sämre förutsättningar med anledning av att deras föräldrar har mindre erfarenhet av det svenska skolsystemet, detta då de själva sannolikt inte har gått i en svensk skola. Språkkunskap och dylika faktorer skulle också kunna tänkas vara en orsak.

5.1.3 Föräldrarnas utbildningsnivå

Den logistiska regressionsmodellen visar att då minst en av föräldrarna har en högskoleutbildning på minst tre år så ökar oddset för att få ett bra betyg med 16 % (9, 22). Resultatet kan anses rimligt och Skolverket har i sin rapport [9] indikerat liknande samband. Föräldrar med hög utbildningsnivå kan sannolikt i högre grad bidra med att lära ut studievana, och till viss del ämneskunskaper, till sina barn. Vidare tror jag att de också kan värdera sitt barns studieresultat högre än föräldrar med lägre utbildningsnivå. Min tankegång är som sådan att den utbildning som dessa föräldrar har anförskaffat sig har sannolikt gett behållning, karriärmässigt eller dylikt. Detta gör i sin tur att de kan se nyttan i sitt barns utbildning på ett tydligare sätt och kommer således sannolikt värdera vikten av ett gott studieresultat högre. Och om elever märker att deras föräldrar bryr sig om deras skolgång och studieresultat kommer de sannolikt också i högre grad bry sig om sitt studieresultat, och således prestera bättre.

5.1.4 Kön

Att vara man innebär att oddset för att uppnå ett bra betyg reduceras med 39 % (36, 42). Skolverket hittar liknande effekter i sin rapport [9] och enligt min uppfattning är att i Sverige är det mer eller mindre allmänt känt att tjejer presterar bättre skolan än killar och vårt resultat är således inte förvånande.

5.1.5 Storstad

Att studera på en skola i en större stad innebär att ditt odds som elev för att uppnå ett gott slutbetyg reduceras med 15 % (10, 20). Detta var för mig personligen lite oväntat då jag tycker mig ha hört att större städer presterar bättre studieresultatmässigt. Jag hittar dock inga tidigare studier gjorda på samma område. Sannolikt varierar detta väldigt mycket med definitioner och gränsdragningar. Vi valde i denna undersökning att bara göra lättaste möjliga distinktion, eftersom vårt övergripande mål var att skapa en enkel modell att tolka. Det är mycket möjligt att vi hade fått ett helt annorlunda resultat om vi exempelvis endast hade tittat på storstäderna (dvs. Stockholm, Göteborg och Malmö) eller definierat innerstadsskolor som en egen grupp. Problemet är dock att det finns väldigt många olika kombinationer i hur man kan definiera gränserna, alldeles för många för att vi skall orka prova runt bland dessa. Jag vet inga specifika argument till varför man borde välja just denna indelning vi valde, bortsett från att den är enkel. Samtidigt vet jag heller inga argument till varför indelningen är olämplig. Det skall också tilläggas att SKL har ändrat definitionerna i sin kommunindelning år 2017 från tidigare år.

5.1.6 Friskola

Att studera på en friskola medför en ökning av oddset för att uppnå ett gott slutbetyg med 80 % (70, 91) enligt vår modell. Det har under de senaste åren flitigt debatterats fram och tillbaka om friskolornas nytta eller onytta. Koncensus verkar vara att friskolor ger ett högre betyg än kommunala, och så är fallet onekligen i vår modell. Skolverket lyfter dock inte fram några liknande effekter.
I debatten verkar det dock som att friskoleelverna sedan presterar sämre, mao. är det alltså frågan om ett s.k. ”överbetyg”. Vi nöjer oss dock med att konstatera ovanstående effekt, men i frågan om kunskapsnivå yttrar vi oss inte. Inte heller här lyckas jag hitta jämförande studier, bortsett från några tidningsartiklar utan specifika källhänvisningar.

5.1.7 Program

Detta är i mina ögon den mest intressanta parameterskattningen; de olika programmen ger väldigt olika resultat. Basnivån har valts till att vara studier inom programmen Samhällskunskap eller Ekonomi, eftersom att dels är detta den största gruppen och dels kan den ses ligga mitt-emellan i ”studie börda”; något mindre än Naturvetenskap- och Teknikprogrammen men något mer än Estet- och Humanistpgorammen.
Oddskvoten för Estet- och Humaniststudenterna säger oss att eleverna här har ett ökat odds om 33 % (24, 43) för att uppnå ett gott studieresultat gentemot om de skulle studera på ett Samhällsvetenskapligt- eller Ekonomiprogram.
Däremot har de elever som studerar på ett Naturvetenskapligt- eller Teknikprogrammet en klar reducering på 67 % (65, 69) av oddset till att uppnå ett gott betyg jämfört med basnivån att studera på ett Samhällsvetenskapligt- eller Ekonomiprogram. Troligtvis är detta en effekt av att kurser inom Naturvetenskap, Matematik och Teknik är de kurser som många anser vara svårast i gymnasiet och resultatet blir därefter.

En aspekt som måste diskuteras i detta avsnitt är behörighet. Vissa kurser behövs för behörighet i högskolan och många kommer komplettera upp med kurser som de nödvändigtvis inte gillar eller är bra på, men med motivationen att de behöver läsa den kursen för att kunna studera vidare på högskola. Till dessa kurser hör t.ex. Matematik 4, och Fysik 2 som krävs för att få komma in på merparten av ingenjörsutbildningarna. Ett rimligt antagande är att elever väljer program efter sitt intresse och sina styrkor, och dessa två faktorer är också ofta korrelerade. Vid tiden för gymnasievalet är dock många ungdomar fortfarande osäkra i vad de vill vill utbilda sig inom och slutligen jobba med, varför många eftersträvar att få en så bred behörighet som möjligt. En teori är alltså att personer i högre grad väljer Naturkunskaps- och Teknikprogrammen (och i viss mån Ekonomiprogrammet) för att skaffa sig en bred behörighet, och inte alltid för att de tycker att programmet nödvändigtvis är roligast eller att det är deras bästa ämnesområde.
För Estet- och Humanistprogrammen tror jag snarare att elever i högre grad väljer för att de har ett intresse inom just detta ämnesområde.
Resultatet som vi får här är inte riktigt jämförbart med Skolverkets rapport [9], eftersom vi har olika populationer och ställer även lite olika frågor.

5.1.8 Utökad studieomfattning

Värdet på oddskvoten visar att personer som läser fler kurser än vad som är obligatoriskt för att erhålla en gymnasiexamen har 91 % (80, 102) ökning av oddset i att erhålla ett bra betyg. Detta är ett föga förvånande resultat - om man läser fler kurser gör man det sannolikt för att man vill och för att man kan dvs. dessa elever är sannooikt i högre grad ambitiösa och presterar således bättre.

5.1.9 Genomsnittligt matematikbetyg i gymnasieskolan

I takt med att det genomsnittliga matematikbetyget i gymnasiet stiger med en enhet kommer oddset för att uppnå ett gott slutbetyg att öka med 49 % (47, 51). Resultatet är väntat och det finns inte så mycket mer att säga om det.

5.2 Förslag till förbättringar

En modell som vi skulle kunna ha använt är en ordinal logistisk regressionsmodell. Denna skiljer alltså på fler nivåer i responsvariabeln och det hade då varit möjligt att räkna ut hur de olika betygsnivåerna förhåller sig till varandra. I takt med att sofistikationen ökar blir också modellen mer komplex och något mer svårtolkad, framförallt för en lekman. Vi har valt att begränsa oss till en binär logistisk regressionsmodell då tolkningen blir avsevärt enklare.
I diskussionen om variabeln *storstad* så nämnde jag att gränserna för indelningarna är något godtyckligt valda, och min uppfattning är att med bättre demografiska kunskaper och något grundligare analys skulle säkerligen denna del av anlysen kunna förbättras eller fördjupas.
För våra befintliga modeller har vi provat att skapa nya variabler, exempelvis en indikatorvariabel för ett sänkt matematikbetyg. Vi jämför helt enkelt en elevs betyg i matematik i grundskolan med betyget i första gymnasiekursen och om det är så att eleven har sänkt sig så indikerar denna variabeln det. Basnivån för Matematik Indikator är alltså att eleven har behållit sitt betyg eller höjt sitt betyg i gymnasiet, men om eleven sänkt sitt betyg från matematik i grundskolan till gymnasiekursens första kurs så kommer denna variabel ge utslag. Tanken är att fånga upp "glädjebetyg" dvs. där eleven eventuellt fått ett oförtjänt högt betyg i grundskolan.
Vi provade även att räkna ut skolstorlek på de olika skolorna baserat på hur många elever som tog examen under aktuell termin och försökte hitta ett samband till slutbetyget.
Denna kreativa approach gav oss dock inte medel för att skapa en bättre modell. Dock finns det säkerligen fler angreppssätt att ta till.

5.3 Sista ord

Jag skulle vilja tacka Mathias Lindholm för att ha tagit sig an att handleda mig i mitt arbete, och för att ha stöttat och uppmuntrat även när arbetsglädjen har sinat. Jag vill även tacka Ola Hössjer för sitt engagemang då han alltid tagit sig tid att diskutera extra runt ämnen efter sina föreläsningar i kursen *Analys av Kategoridata*. Ett stort tack vill jag rikta till Sven Sundin på Skolverket som dels gav mig tillgång till datamaterialet, men också för att ha varit mycket hjälpsam vid frågor och tillmötesgående vid olika anpassningar i datamaterialet.

6 Appendix

6.1 Utskrifter från linjär regression

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.918002   0.066914  73.498  < 2e-16 ***
meritvarde   0.482000   0.003795 127.008  < 2e-16 ***
matgr       -0.076796   0.002610 -29.427  < 2e-16 ***
utlbakg1    -0.206420   0.020983  -9.837  < 2e-16 ***
utlbakg2    -0.120710   0.028770  -4.196 2.73e-05 ***
foruniva2    0.047894   0.044919   1.066   0.2863
foruniva3    0.097806   0.045773   2.137   0.0326 *
foruniva4    0.197872   0.044938   4.403 1.07e-05 ***
kon1        -0.286511   0.013382 -21.410  < 2e-16 ***
storstad1   -0.124407   0.014583  -8.531  < 2e-16 ***
friskola1    0.406291   0.014029  28.960  < 2e-16 ***
programES    0.271208   0.026826  10.110  < 2e-16 ***
programHU    0.267593   0.050507   5.298 1.17e-07 ***
programNV   -0.983495   0.025246 -38.957  < 2e-16 ***
programSA    0.030418   0.019747   1.540   0.1235
programTE   -0.688402   0.026557 -25.922  < 2e-16 ***
studieomf1   0.409468   0.013690  29.909  < 2e-16 ***
matsnitt     0.303957   0.002525 120.385  < 2e-16 ***
matkurser    0.120804   0.011291  10.699  < 2e-16 ***
--
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.311 on 47009 degrees of freedom
Multiple R-squared:  0.665, Adjusted R-squared:  0.6648
F-statistic:  5183 on 18 and 47009 DF,  p-value: < 2.2e-16
```

```
             Df Sum of Sq    RSS   AIC
<none>                    80781 25480
- storstad    1     125.1  80906 25551
- utlbakg     2     181.1  80962 25582
- foruniva    3     201.3  80983 25591
- matkurser   1     196.7  80978 25593
- kon         1     787.7  81569 25935
- friskola    1    1441.2  82222 26310
- matgr       1    1488.1  82269 26337
- studieomf   1    1537.2  82318 26365
- program     5    2933.4  83715 27148
- matsnitt    1   24904.4 105686 38116
- meritvarde  1   27719.8 108501 39352
```

```
               GVIF Df GVIF^(1/(2*Df))
meritvarde 2.435775  1        1.560697
matgr      2.714543  1        1.647587
utlbakg    1.140978  2        1.033521
foruniva   1.148917  3        1.023406
kon        1.210820  1        1.100373
storstad   1.108027  1        1.052628
friskola   1.085172  1        1.041716
program    5.240236  5        1.180144
studieomf  1.068867  1        1.033860
matsnitt   2.192681  1        1.480771
matkurser  5.013765  1        2.239144
```

6.2 Utskrifter för logistisk regression

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -13.421639   0.172326 -77.885  < 2e-16 ***
meritvarde    0.614335   0.008972  68.471  < 2e-16 ***
matgr        -0.102324   0.005444 -18.796  < 2e-16 ***
utlbakg1     -0.288559   0.044141  -6.537 6.27e-11 ***
utlbakg2     -0.163389   0.062230  -2.626  0.00865 **
foruniva2     0.064680   0.096887   0.668  0.50440
foruniva3     0.111569   0.098640   1.131  0.25802
foruniva4     0.263893   0.096820   2.726  0.00642 **
kon1         -0.495988   0.028634 -17.322  < 2e-16 ***
storstad1    -0.158844   0.031175  -5.095 3.48e-07 ***
friskola1     0.571842   0.030989  18.453  < 2e-16 ***
programES     0.472544   0.057400   8.232  < 2e-16 ***
programHU     0.302097   0.107807   2.802  0.00508 **
programNV    -1.441509   0.055169 -26.129  < 2e-16 ***
programSA     0.037024   0.042047   0.881  0.37857
programTE    -0.949626   0.057440 -16.533  < 2e-16 ***
studieomf1    0.605248   0.030141  20.081  < 2e-16 ***
matsnitt      0.439261   0.006694  65.618  < 2e-16 ***
matkurser     0.179677   0.023997   7.487 7.03e-14 ***
--
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 64601  on 47027  degrees of freedom
Residual deviance: 36841  on 47009  degrees of freedom
AIC: 36879

Number of Fisher Scoring iterations: 5
```

```
            Df Deviance   AIC
<none>           36841 36879
- storstad   1    36867 36903
- foruniva   3    36889 36921
- utlbakg    2    36887 36921
- matkurser  1    36897 36933
- kon        1    37143 37179
- friskola   1    37188 37224
- matgr      1    37202 37238
- studieomf  1    37251 37287
- program    5    37624 37652
- matsnitt   1    42563 42599
```

```
               GVIF Df GVIF^(1/(2*Df))
meritvarde 1.679813  1        1.296076
matgr      2.135079  1        1.461191
utlbakg    1.136169  2        1.032430
foruniva   1.100639  3        1.016110
kon        1.198965  1        1.094973
storstad   1.108981  1        1.053082
friskola   1.108082  1        1.052655
program    5.200479  5        1.179246
studieomf  1.051123  1        1.025243
matsnitt   1.578506  1        1.256386
matkurser  4.743525  1        2.177963
```

I tabellerna nedan finns parameterskattningarna för den logistiska modellen med ett genomsnittligt slutbetyg D eller bättre, respektive ett genomsnittligt slutbetyg på B eller bättre.

$slutbetyg \geq D$	Estimat	Std. avvikelse
α	-8.67	0.148
$meritvarde$	0.56	0.010
$utlbakg = 1$	-0.28	0.044
$foruniva = 1$	0.14	0.035
$kon = 1$	-0.54	0.037
$storstad = 1$	-0.24	0.041
$friskola = 1$	0.43	0.040
$program = ES/HU$	0.37	0.051
$program = NV/TE$	-0.61	0.043
$studieomf = 1$	0.63	0.044
$matsnitt$	0.28	0.007

$slutbetyg \geq B$	Estimat	Std. avvikelse
α	-17.67	0.199
$meritvarde$	0.61	0.011
$utlbakg = 1$	-0.11	0.052
$foruniva = 1$	0.14	0.038
$kon = 1$	-0.22	0.035
$storstad = 1$	-0.10	0.040
$friskola = 1$	0.72	0.037
$program = ES/HU$	0.02	0.050
$program = NV/TE$	-1.41	0.041
$studieomf = 1$	0.58	0.034
$matsnitt$	0.44	0.007

Referenser

[1] Andersson, Patrik & Tyrcha, Joanna
Notes in Econometrics
Stockholm, 2012.

[2] Statistiska Centralbyrån
Befolkningens Utbildning 2013
http://www.scb.se/sv_/Hitta-statistik/Statistik-efter-amne/Utbildning-och-forskning/Befolkningens-utbildning/Befolkningens-utbildning/9568/9575/Behallare-for-Press/372838/
Besökt 2015-12-07

[3] Sveriges Kommuner och Landsting
Kommungruppsindelning 2017
https://skl.se/tjanster/kommunerlandsting/faktakommunerochlandsting/kommungruppsindelning.2051.html
Besökt 2017-03-05

[4] Alm, Sven Erick & Britton, Tom.
Stokastik
Liber AB, 2012. Första upplagan, andra tryckningen.
ISBN 978-91-47-05351-3.

[5] Agresti, Alan.
Categorical Data Analysis, 3rd edition.
Wiley-Interscience, 2013.

[6] Sundberg, Rolf.
Kompendium i Lineara Statistiska Modeller
Matematisk Statistik, Stockholms Universitet., Oktober 2014.

[7] Fox, John & Monette, Georges.
Generalized Collinearity Diagnostics
Journal of the American Statistical Association, Vol. 87, No. 417 (Mar., 1992) pp. 178-183
http://www.jstor.org/stable/2290467
Besökt 2017-05-20

[8] *R documentation*
https://stat.ethz.ch/R-manual/R-devel/library/stats/html/glm.html
Besökt 2017-05-10

[9] Skolverket
PM - Betyg och studieresultat i gymnasieskolan 2013/2014
https://www.skolverket.se/publikationer?id=3368
Besökt 2017-04-20

[10] Gut, Allan.
An Intermediate Course in Probability, Second Edition
Springer, 2009.